

# The Determination of Sample Size in a Bayesian Estimation of Population Proportions: How and Why to Do It in a Regression Framework

R. A. Weitzman

Naval Postgraduate School, Monterey, USA

**Email address:**

[ronweitzman@redshift.com](mailto:ronweitzman@redshift.com)

**To cite this article:**

R. A. Weitzman. The Determination of Sample Size in a Bayesian Estimation of Population Proportions: How and Why to Do it in a Regression Framework. *American Journal of Theoretical and Applied Statistics*. Vol. 11, No. 1, 2022, pp. 13-18.

doi: 10.11648/j.ajtas.20221101.12

**Received:** December 28, 2021; **Accepted:** January 17, 2022; **Published:** January 21, 2022

---

**Abstract:** Principally to reduce the cost by reducing the sample size required to conduct survey research, this article presents and illustrates the use of a method to determine the sample sizes required to obtain Bayesian estimates of population proportions with specified margins of error. The development proceeds within a regression framework derived from mental test theory. Specifically, building on prior work, the development presented here enables a researcher to conduct a survey to obtain pure Bayes estimates of the proportion of all members of a defined population choosing each one of a number of mutually exclusive and exhaustive options or falling into each one of a number of mutually exclusive and exhaustive categories, including two. The regression framework not only provides useful insight into Bayesian and classical statistics but also enables the development to proceed without explicit reference to the differing parent distributions of the sample and population proportions, both being asymptotically normal. In addition to the sample-size advantage, which is substantial, this article identifies other practical advantages that Bayesian has over classical estimation of population proportions and, in a somewhat in-depth comparison of the two, discusses other reasons a Bayesian method may be a powerful substitute for the classical method of estimating population proportions via independent random sampling.

**Keywords:** Population Proportion, Survey Research, Pure Bayes Estimate, Regression Model, Standard Error of Estimate, Sample Size

---

## 1. Introduction

Survey research requires large samples. This requirement is both an advantage and a disadvantage, at least in the estimation of population proportions. The advantage is that the frequency distribution of a sample proportion or, in the case of Bayesian estimation, a population proportion tends for all practical purposes to be normal. The disadvantage is that the high cost of large samples limits the use of survey research. This article has three purposes. The first is to present a method that substantially reduces the sample size and cost required to conduct survey research. The second is to show that the Bayesian estimation method described here satisfies the first purpose. The third purpose is to demonstrate that use of a regression framework makes development of the method possible without explicit or necessary involvement of

the parent distribution of a sample or a population proportion, each being asymptotically normal.

Although Bayesian estimation generally tends to require a smaller sample size ( $n$ ) to achieve a specified margin of error ( $m$ ) than classical estimation, the determination of that sample size can be a problem. In the estimation of population proportions, only classical estimation with independent random sampling has a simple formula for determining  $n$ :  $n = 0.96/m^2$  for an error probability equal to .05. In the absence of such a formula, this article will describe a patented process involving simulation for determining values of  $n$  needed to achieve specified values of  $m$  in a Bayesian estimation of population proportions. (Commercial use of this process or its results requires a license.) This article will also provide results and illustrate the applicability of the process for cases involving two or more options ( $K$ ). Following a brief review of the literature and some

preliminary development, the presentation will begin with the case  $K = 2$  and then address the general case  $K \geq 2$ . For some readers, now may be a good time to read Section 6.

## 2. Background

### 2.1. Prior Work

In addition to the classical method involving independent random sampling, a number of Bayesian-like methods exist for estimating population from sample means or proportions. About 60 years ago, James and Stein [7] derived estimators of population means that are more efficient than corresponding classical estimators by using a linear combination of the mean of each individual sample and the overall mean of the sample aggregated with two or more other samples from possibly different populations. Being within the 0-1 interval, the weight applied to each individual sample mean has been called a shrinkage coefficient.

Commenting on the empirical Bayes treatment of James-Stein estimators by Efron and Morris [3] and Morris [9], Stigler [12, 13] showed that the shrinkage coefficient was an estimator of the squared correlation coefficient in the regression of population on sample means. Fienberg and Holland [4] extended the empirical Bayes treatment of James-Stein estimators to the estimation of population proportions, with the expected increase in efficiency. Then, Weitzman [14] showed that the shrinkage coefficient in a Bayesian estimation of population proportions is the squared correlation between sample ( $P$ ) and population ( $\pi$ ) proportions and, with no assumptions other than independent random sampling, developed the following classical estimator for it:

$$\hat{\rho}_{\pi P}^2 = 1 - \frac{K(1 - \sum_{k=1}^K P_k^2)}{(n-1)(K \sum_{k=1}^K P_k^2 - 1)} \quad (1)$$

Noteworthy in this formula is that it can yield a value between 0 and 1 even when  $K = 2$  because  $P$  and  $\pi$  have equal means ( $1/K$ ).

In the regression of the observed proportion  $P$  toward the mean,  $1/K$ , the squared correlation  $\hat{\rho}_{\pi P}^2$  resembles but is not identical to the shrinkage coefficient  $w$  in Fienberg and Holland [4] and  $1 - B$  in Efron and Morris [3] and Morris [9]. Different from the development using  $w$  or  $1 - B$ , which is empirical Bayes, the development described here, which builds on Weitzman [14], is pure Bayes and, as the comparison in Section 3 shows, is the more efficient of the two for the estimation of population proportions.

Other than Fienberg and Holland [4] and Weitzman [14], the literature shows almost no connection between “Bayes or Bayesian estimation” and “estimation of proportions.” Articles on the second of those two subjects have focused on the possible need to adjust samples, for example by weighting sampled responses as in Pfeiffermann [11], to assure a demographic match between samples and populations or other such sampling problems, e.g., Healy [6] and Alvarez et al. [1]. A literature search has identified only Novick et al. [10] as making the connection, but there the

focus was on a single option in multiple groups rather than multiple options in a single group, as here.

### 2.2. Regression Framework

The regression framework underlying the development here derives from mental test theory. In Gulliksen's classical *Theory of Mental Tests* [5], the formula  $X = T + E$  describes the relationship between true ( $T$ ) and observed ( $X$ ) scores. That being the case, the regression equation

$$X = \left( \frac{S_X}{S_T} \right) r_{XT} (T - \bar{T}) + \bar{X} + E \quad (2)$$

implies that  $r_{XT} = S_T / S_X$  and  $\bar{X} = \bar{T}$ . The relationship  $r_{XT} = S_T / S_X$  is central in mental test theory. The standard error of measurement in this theory ( $S_X \sqrt{1 - r_{XT}^2}$ ) is the standard deviation of the error component ( $E$ ) of  $X$ .

Prior to the publication of Gulliksen's *Theory of Mental Tests*, Kelley [8] presented a regression formula for estimating an examinee's true test score from the examinee's observed test score, the two means again being equal:

$$\hat{T} = r_{XT}^2 X + (1 - r_{XT}^2) \bar{X} \quad (3)$$

The standard error of estimate in Kelley's test theory ( $S_T \sqrt{1 - r_{XT}^2}$ ) is the standard deviation of the error component of  $T$ .

Although regression underlies both the Kelley and Gulliksen models, the difference between them stems from the *direction* of the regression. Although so many researchers appear to be unaware of the difference or to discount it as irrelevant, a problem motivating articles like Charter [2] to clarify the issue, the difference is in fact profound. As Weitzman [15] has shown, in contrast to Gulliksen's classical estimator, Kelley's is a Bayesian estimator. This article builds on that difference.

## 3. Comparison of Sample Sizes in Different Estimation Methods

### 3.1. Pure Bayes vs. Empirical Bayes Estimation

In reference to the formula for the error variance of  $\pi$  having a Dirichlet distribution with  $\hat{\pi}_k$  designating a fixed sample estimate for option  $k$  when  $K \geq 2$ ,

$$\sigma_{\pi - \hat{\pi}}^2 = \frac{\hat{\pi}_k(1 - \hat{\pi}_k)}{\tau + n + 1} \quad (4)$$

Weitzman [14] compared the formula for  $\tau$  developed by Fienberg and Holland [4] in empirical Bayes estimation,

$$\tau = \frac{K(1 - \sum_{k=1}^K P_k^2)}{K \sum_{k=1}^K P_k^2 - 1} \quad (5)$$

with the formula developed by Weitzman after first showing

that as a function of  $\rho_{\pi P}^2$  in pure Bayes estimation  $\tau = n(1 - \rho_{\pi P}^2)/\rho_{\pi P}^2$  and then substituting for  $\rho_{\pi P}^2$  from Equation (1),

$$\tau = \frac{K(1 - \sum_{k=1}^K P_k^2)}{K \sum_{k=1}^K P_k^2 - 1 - \frac{K-1}{n}} \quad (6)$$

Since for any value of  $n > 0$  the  $\tau$  of Equation 6 will be larger than the  $\tau$  of Equation 5, the error variance of Equation 4 will correspondingly be smaller for the Weitzman than for the Fienberg-Holland method. That is an expected difference because Weitzman's is a pure Bayes method in which only populations can vary whereas in the empirical Bayes method of Fienberg and Holland both samples and populations can vary.

### 3.2. Pure Bayes vs. Classical Estimation

The regression framework used here, with  $\rho_{\pi P}^2$  from Equation 1, permits a similar comparison confirming the efficiency advantage of pure Bayes over classical estimation. In this framework (with  $\epsilon$  representing error), the classical model ( $P = \pi + \epsilon$ ) has the standard error of measurement

$$\sigma_{P-\pi} = S_P \sqrt{1 - \rho_{\pi P}^2} \quad (7)$$

and the pure Bayes model ( $\pi = \hat{\pi} + \epsilon$ ) has the standard error of estimate

$$\sigma_{\pi-\hat{\pi}} = \sigma_{\pi} \sqrt{1 - \rho_{\pi P}^2} \quad (8)$$

In the classical model,  $P$  is regressed on  $\pi$  with  $\rho_{\pi P} = \sigma_{\pi}/S_P$  whereas, in the pure Bayes model,  $\pi$  is regressed on  $P$ , with

$$\hat{\pi} = \rho_{\pi P}^2 P + (1 - \rho_{\pi P}^2) \mu \quad (9)$$

where  $\mu = 1/K$  for both  $P$  and  $\pi$ . Since  $\sigma_{\pi} = \rho_{\pi P} S_P$ , Equations 7 and 8 show that the error variance of a pure Bayes estimate will be smaller than the error variance of a classical estimate for equal values of  $\rho_{\pi P}$  ( $0 < \rho_{\pi P} < 1$ ). The practical implication of this difference in estimation efficiency is that a pure Bayes estimate will require a smaller sample than a classical estimate to achieve error variances of equal size.

### 3.3. Extra Sample Size Reduction for Pure Bayes Due to Concentration of the Error Probability in a Single Tail

This lower- $n$  advantage over classical estimation, shared by pure Bayes and empirical Bayes, is even greater for pure Bayes because of a difference between it and both of the other two methods of estimation. In both those methods, the value of a sample statistic can be on either side of the population parameter value of concern (classical) or its multiple-sample surrogate (empirical Bayes) with the error probability (.05) being divided equally between the two tails (.025 in each tail) of the sample statistic's probability distribution. Contrariwise, the value of a pure Bayes sample statistic (here,  $\hat{\pi}$ ), being fixed, can have the population parameter value of concern (here,  $1/K$ ) on only one side of it, with the error probability (.05) being entirely on that side of the parameter's likelihood distribution centered at the

value of the sample statistic. Because a distribution that cuts off a .05 tail on one side must be wider than a distribution that cuts off a .025 tail at the same point on that side and because a wider distribution requires a smaller sample size, a pure Bayes sample statistic will require, for equal margins of error, an even smaller sample than it would otherwise in comparison with either a classical or an empirical Bayes estimate.

From here on, as previously in this article, the term "Bayesian" without the qualifier "pure" will refer specifically to the pure Bayes method of estimation.

## 4. The Case of $K = 2$

The Bayesian estimation of a population proportion described here requires the following steps (in the language of its patent):

Step 1. Specify a desired margin of error for a poll comprising a plurality of options;

Step 2. Determine a sample size needed to achieve the desired margin of error by simulating samples of different sizes, each yielding a margin of error as a function of not only the sample size but also the desired margin of error and the number of the plurality of options, the determined sample size being the one yielding the desired margin of error;

Step 3. Poll an independent random sample of the determined sample size;

Step 4. Convert a proportion of individuals choosing a particular one of the plurality of options in the poll into a Bayesian point estimate of the population proportion choosing the particular one of the plurality of options by a simple regression of the Bayesian point estimate on the polled proportion; and

Step 5. Present the results of the poll, wherein the sample size is smaller than a sample size required to achieve the desired margin of error without the simulation and the conversion.

Focusing on Step 2, the development of a process to determine the sample size required to achieve a specified margin of error will begin with the following version of the formula for  $\hat{\rho}_{\pi P}^2$ :

$$\hat{\rho}_{\pi P}^2 = \frac{\left(\frac{n}{n-1}\right) S_P^2 - \left(\frac{K-1}{n-1}\right) \left(\frac{1}{K}\right)^2}{S_P^2} \quad (10)$$

presented by Weitzman [14] who used this formula to show that in classical, estimation  $\hat{\rho}_{\pi P}^2$  is an unbiased estimate of  $\rho_{\pi P}^2$ , the numerator  $S_P^2$  being variable and the denominator  $S_P^2$  being fixed. Conversely, in pure Bayes estimation the numerator  $S_P^2$  and the denominator  $S_P^2$  are both fixed. In pure Bayes estimation, the roles of  $\pi$  and  $\hat{\pi}$  are reversed with  $\hat{\pi}$  effectively (but not literally) giving its hat to  $\pi$  so that what was variable is now fixed and what was fixed is now variable. The implication here is that, as a pure Bayes estimator, the  $\hat{\rho}_{\pi P}^2$  of Equations 1 and 10 loses its hat so that as a pure Bayes estimator Equation 10, particularly, becomes

$$\rho_{\pi P}^2 = \frac{\left(\frac{n}{n-1}\right) S_P^2 - \left(\frac{K-1}{n-1}\right) \left(\frac{1}{K}\right)^2}{S_P^2} \quad (11)$$

with  $S_P^2 = \sigma_{\pi}^2 / \rho_{\pi P}^4$  since, from Equation 9,  $\sigma_{\pi}^2 = \rho_{\pi P}^4 S_P^2$ .

For  $K = 2$  and  $\hat{\pi} > .5$ , the value of  $n$  needed to achieve a margin of error equal to  $m$  is the one for which  $\hat{\pi} = .5 + m$  is the center of the likelihood distribution of  $\pi$ . This plus Equations 12 below is all the information needed to make a substitution for  $S_P^2$  in Equation 11 and solve for  $\rho_{\pi P}^2$  as a function of  $n$  and  $m$  with  $K = 2$ .

Since, with  $K = 2$ ,

$$\sigma_{\pi}^2 = \frac{\hat{\pi}_1^2 + \hat{\pi}_3^2}{2} - \frac{1}{4} = \frac{(.5+m)^2 + (.5-m)^2}{2} - \frac{1}{4} = m^2 \quad (12)$$

substitution of  $m^2 / \rho_{\pi P}^4$  for  $S_P^2$  in Equation 11 and solving for  $\rho_{\pi P}^2$  yields

$$\rho_{\pi P}^2 = 2 \left( \sqrt{((n-1)m^2)^2 + nm^2} - (n-1)m^2 \right) \quad (13)$$

The relationship between  $\sigma_{\pi}^2$  and  $\sigma_{\pi}^2$  is notable: Substitution of  $S_P^2 = \sigma_{\pi}^2 / \rho_{\pi P}^4$  for  $S_P^2$  in  $\sigma_{\pi}^2 = \rho_{\pi P}^4 S_P^2$  and solving for  $\sigma_{\pi}^2$  shows that  $\sigma_{\pi}^2 = \sigma_{\pi}^2 / \rho_{\pi P}^4$  or, with  $m^2$  for  $\sigma_{\pi}^2$ ,  $\sigma_{\pi}^2 = m^2 / \rho_{\pi P}^4$ .

Completion of Step 2 requires the computation of  $1.645\sigma_{\pi-\hat{\pi}}$  using either Equation 4 with  $\hat{\pi}_1 = .5 + m$  and  $\tau = n(1 - \rho_{\pi P}^2) / \rho_{\pi P}^2$  or Equation 8 with  $\sigma_{\pi} = |m / \rho_{\pi P}|$ , together with  $\rho_{\pi P}^2$  determined from Equation 13, for successive values of  $n$  until reaching the value of  $n$  for which  $1.645\sigma_{\pi-\hat{\pi}}$  is equal to  $m$ , which is the value of  $n$  needed to achieve a margin of error equal to  $m$  in Bayesian estimation of a population proportion with  $K = 2$ .

Table 1 shows values of  $n$  needed for achieving different values of  $m$  in one-sided and two-sided Bayesian estimation and in classical, two-sided estimation. The values of  $n$  for  $m = .03$  are of especial interest in their indication of the great extent to which one-sided Bayesian estimation can be more efficient than two-sided Bayesian and classical estimation: 525, 825, and 1,067, in that order.

**Table 1.** Bayesian and Classical Sample Sizes for Different Margins of Error When  $K = 2$ .

Error Margin	Bayesian (one-sided)	Bayesian (two-sided)	Classical (two-sided)
.03	525	825	1,067
.035	400	600	784
.04	300	475	600
.045	250	375	474
.05	200	300	384

NOTE: Bayesian sample sizes were determined in simulations with increments of 25.

In addition to its requirement of about half the sample size required by classical estimation to achieve a specified margin of error, Bayesian estimation with  $K = 2$  can demonstrate its superior power in other ways, as illustrated by the following fictitious examples.

#### 4.1. Example 1. Comparison of Classical and Bayesian Estimation

A water activist group which is planning to conduct a

referendum to compel a city council to purchase and operate its local water utility has a budget of \$18,000 to conduct a poll to determine voter support for the purchase. A pollster who agrees to conduct the poll at that price tells the group it can be done with a sample of 600, which would have a classical margin of error equal to .04. Conducting the poll, the pollster finds that the proportion in favor of the referendum,  $P_1$ , is equal to .53 (318 in favor) and reports to the group that result would be outside a .03 margin of error but not, unfortunately, outside the .04 one of their poll. Sharing the group's unhappiness with the report, the pollster recalls hearing about a possible Bayesian analysis of the data that might produce a more favorable outcome. Conducting that analysis, the pollster computes  $\hat{\pi}_1$  from Equation 9 with  $\mu = .5$  (the value of  $1/K$ ) after determining  $\rho_{\pi P}^2$  from Equation 1 and the actual Bayesian margin of error as  $1.645\sigma_{\pi-\hat{\pi}}$  from Equation 4 using  $\tau = n(1 - \rho_{\pi P}^2) / \rho_{\pi P}^2$ , with the following results:  $\rho_{\pi P}^2 = .54$  and  $\hat{\pi}_1 = .52$ , which this time is a favorable outcome having an actual Bayesian margin of error equal to .02, all to two decimal places. The group can now have increased confidence that its plan might succeed.

#### 4.2. Example 2. Use of Performance Assessment in Choosing a Textbook for a Course

In considering the possible choice of a new textbook for a high school science course, a state education department decides to compare the performance on a state final examination by two groups of students, the first group using the new and the second the current textbook, a huge supply of which is stored in a state warehouse. Seeking a wide performance margin between the groups to justify a change of textbooks, which would be costly, the department chooses a margin of error equal to .05. According to Table 1, that choice requires a sample size of 200. So, the department assigns 200 students to each group in a random selection of high school classes throughout the state. At the end of the study, which takes a year, the department lines up all 400 student scores on the final examination in order from highest to lowest, with 112 members of the first group and 88 members of the second group performing above the median score:  $P_1 = .56$  and  $P_2 = .44$ . Using Equation 1 with  $K = 2$  and  $n = 200$ , the department determines that  $\rho_{\pi P}^2 = .66$  and then, using Equation 9 that  $\hat{\pi}_1 = .54$  and  $\hat{\pi}_2 = .46$ . Because .50 is within the .05 margin of error for either  $\hat{\pi}$  value, the study fails to show that the new textbook is sufficiently more effective than the current one to warrant a change of textbooks.

Although the study required 400 students to have a value of  $n$  equal to 200, a classical t-test of the difference between the means of the two groups would also require double the number of members in each group. Despite that equivalency, the department chose to compare percentages rather than means because, of the two, only the comparison of percentages involves explicit use of the relationship between  $m$  and  $n$  needed to meet the study's objectives. Use of a Bayesian rather than a classical method also reduced the value of  $n$  and the cost of the study by about half.

## 5. The Case of $K \geq 2$

For  $K \geq 2$  and  $\hat{\pi} > 1/K$ , the value of  $n$  needed to achieve a margin of error equal to  $m$  is the one for which  $\hat{\pi} = .5 + (m + 1/K - .5)$  is the center of the likelihood distribution of  $\hat{\pi}$ . If  $K = 4$  and  $m = .03$ , for example,  $n$  would have to have the value for which  $\hat{\pi} = .28$  is the center of the distribution of  $\pi$ . If  $M = m + 1/K - .5$ , then, for  $K \geq 2$ ,  $\hat{\pi} = .5 + M$  would place  $1/K$  outside the margin of error. With the substitution of  $M$  for  $m$ , the equations used to determine  $n$  when  $K = 2$  can now be used to determine  $n$  when  $K > 2$  for specified values of  $m$ . Table 2 shows the results of using this procedure for several values of  $K$  and  $m$ .

**Table 2.** Bayesian Sample Size for Margin of Error ( $m$ ) and Number of Options ( $K$ ).

	K				Classical Sample Size
	2	3	4	5	
.03	525	675	600	525	1,067
.035	400	500	450	400	784
.04	300	375	350	300	600
.045	250	300	275	250	474
.05	200	250	225	200	384

Because  $n$  decreases as  $K$  increases when  $K > 2$ , the relative values of  $n$  for  $K \leq 3$  may come as a surprise. Why for any value of  $m$  is  $n$  larger for  $K = 3$  than it is for  $K = 2$ ? The answer involves values of  $\rho_{\pi P}^2$ . Values of  $M^2$  and, correspondingly, values of  $\rho_{\pi P}^2$  are considerably larger when  $K = 3$  than when  $K = 2$ , and increasingly large values of  $\rho_{\pi P}^2$  bring Bayesian closer to classical estimation and its  $n$  requirements.

The following example, though fictitious, illustrates not only the sample-size advantage over classical estimation but also the unique usefulness of solid Bayesian estimates of population proportions when  $K > 2$ , so solid that the estimated and actual population proportions are all but indistinguishable.

### 5.1. Example Selecting Groups for Targeted Messages

A federal government agency trying to encourage adults who are not vaccinated against Covid-19 to get vaccinated believes it could strengthen its message with increased precision if it knew the likelihood of hospitalization for Covid-19 within select co-morbidity groups, particularly obese adults who do not smoke (Group 1), smokers who are not obese (Group 2), and smokers who are obese (Group 3), as well as adults who are not obese and who do not smoke (Group 4), constituting the following proportions of the United States adult population:  $Q_1 = .42$ ,  $Q_2 = .14$ ,  $Q_3 = .05$ , and  $Q_4 = .39$ . Along with these numbers, to get the results they are looking for with a margin of error of .03 in Bayesian estimation, the agency's researchers take a national random sample of 600 (out of a total of over 6,000) hospitals requesting each to identify the first Covid-19 adult patient admitted to the hospital after receiving the request by the patient's membership in one of the four groups, with the following results:  $P_1 = .32$ ,  $P_2 = .20$ ,  $P_3 = .41$ , and  $P_4 = .07$ . To determine the Bayesian counterparts of these results, the researchers compute  $\rho_{\pi P}^2$  from Equation 1 and then, from Equation 9,  $\hat{\pi}_1 = .32$ ,  $\hat{\pi}_2 = .20$ ,

$\hat{\pi}_3 = .41$ , and  $\hat{\pi}_4 = .07$ , with  $\rho_{\pi P}^2 > .99$ , all to two decimal places. With that extremely high correlation, these are indeed solid estimates. If this were an actual study, it would show, among other things that obese people who smoke account for 41 percent of adults hospitalized for Covid-19 though they comprise only 5 percent of the adult population, some of whom have been vaccinated or have previously contracted the virus. Using this method of estimation, the agency could sharpen the targeting of its message instead of scattershooting it to the entire adult population.

### 5.2. Two Possible Sample Sizes for Each Combination of $K$ and $m$ Values When $K > 2$

When  $K > 2$ , a desired margin of error may actually be achieved with either of two sample sizes, depending on the intent of the study. The sample size will be larger if the intent is to identify options having proportions above average by at least the margin of error, and correspondingly lower otherwise. The reason is that  $\hat{\pi} = .5 + M$  is two-times- $|m|$  closer to zero for the below-average option, when  $m < 0$  in the formula for  $M$ , than for the above-average option. For that reason, the sample sizes in Table 2 are conservative. If  $K = 4$  and  $m = .03$ , for example, the larger  $n$  would be equal to 580, but the smaller one, with  $m = -.03$ , would be considerably smaller: 499. The larger  $n$  differs from—and is more accurate than—the value in Table 2 because it was determined in simulation with increments of 1 rather than 25, as was the smaller  $n$ . Researchers who use this simulation method should do likewise.

## 6. Discussion

Rather than being only an adjunct to the determination of sample sizes, as is the case in some Bayesian literature, the Bayesian method described here is a powerful substitute for the classical method of estimating population proportions via independent random sampling. Both classical and Bayesian estimation of population proportions involving independent random sampling require taking a single sample from a single population. The essential difference between the two estimation methods is that in classical estimation the sample is considered to be one of many that the population can produce whereas in Bayesian estimation the population is considered to be one of many that can produce the sample. In classical estimation, the population is fixed and the sample can vary; in Bayesian estimation, the sample is fixed and the population can vary. Although that difference alone may or may not justify the choice of one over the other, their difference in sampling power favors the Bayesian method, at least in the realm of independent random sampling.

Why might the difference between population and sample variation affect the choice between the two methods? In the classical case of sample variation, the logic is deductive: If A (the null hypothesis is true), then B (the sample estimate will be within the specified margin of error). The conclusion is that A is not true (the null hypothesis is false) if B is not true, with no conclusion otherwise. In the Bayesian case of population variation, the logic is inductive: The sample estimate (A) is

empirically the most likely value of its population counterpart, and any value on either side of A is increasingly less likely the farther it is away from A. If the population counterpart of interest (B) is so far away that it is outside the margin of error, then the conclusion is that B is too unlikely to be true, with no conclusion otherwise. Both methods are logical.

Some may prefer the deductive method because they “know” that the population is fixed since they take samples from lists of addresses or telephone numbers that define it. What they know is that the lists are fixed, but specific numerical descriptions of the listed individuals vary in their likelihood, and that is the variation that is meant by the term population variation. The choice between the two methods conceptually really boils down to a preference of one version of logic over another. Because of its greater power, the choice empirically favors the Bayesian method.

Aside from those conceptual and practical considerations, together with the particular Bayesian method described here, this article demonstrates how a regression framework can provide an insightful and useful characterization of classical and Bayesian statistics.

## 7. Conclusion

Margins of error are important in all estimations, particularly so in the estimation of population proportions. Because sample sizes affect margins of error, knowing how to determine sample sizes from margins of error is important. That knowledge is especially important in Bayesian estimation because, at least in the case of independent random sampling, Bayesian estimates tend to require smaller samples than classical estimates having equal margins of error. For that reason, finding out how to determine sample sizes from margins of error in a Bayesian estimation of population proportions was the principal motivation for the research reported in this article.

## References

- [1] Alvarez, E., Arcos, A., Gonzalez, S., Munoz, J. F., & Rueda, M. (2013). Estimating population proportions in the presence of missing data, *Journal of Computational and Applied Mathematics*, 237 (1), 470-476.
- [2] Charter, Richard A. (1992). Revisiting the Standard Errors of Measurement, Estimate, and Prediction and Their Application to Test Scores, *Perception and Motor Skills*, 82 (3c), 1139-1144.
- [3] Efron, B., & Morris, C. (1973). Stein's Estimation Rule and its Competitors—an Empirical Bayes Approach. *Journal of the American Statistical Association*, 68, 117-130.
- [4] Fienberg, S. E., & Holland, P. W. (1973). Simultaneous Estimation of Multinomial Cell Probabilities. *Journal of the American Statistical Association*, 68, 683-691.
- [5] Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley.
- [6] Healy, J. D. (1981). The Effects of Misclassification Error on the Estimation of Several Population Proportions, *The Bell System Technical Journal*, 60 (5), 697-705.
- [7] James, W., & Stein, C. M. (1961). Estimation with Quadratic Loss Function. *Proceedings of the 4th Berkeley Symposium*, 1, 361-379.
- [8] Kelley, T. L. (1947). *Fundamental Statistics*. Cambridge, MA: Harvard University Press.
- [9] Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, 78, 47-59.
- [10] Novick, M. R., Lewis, C., & Jackson, P. H. (1973). The Estimation of Proportions in  $m$  Groups, *Psychometrika*, 38, 19-46.
- [11] Pfeiffermann, Danny (1993). The Role of Sampling Weights When Modeling Survey Data, *International Statistical Review / Revue internationale de Statistique*, 317-337.
- [12] Stigler, S. M. (1983). Comment on C. Morris, Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, 78, 62-63.
- [13] Stigler, S. M. (1990). The 1988 Neyman Memorial Lecture: a Galtonian Perspective on Shrinkage Estimators. *Statistical Science*, 5, 147-155.
- [14] Weitzman, R. A. (2006). Population-Sample Regression in the Estimation of Population Proportions. *Journal of Educational and Behavioral Statistics*, 31, 413-43.
- [15] Weitzman, R. A. (2009). Fitting the Rasch Model to Account for Variation in Item Discrimination, *Educational and Psychological Measurement*, 69, 216-231.