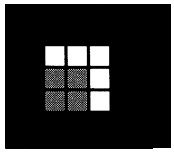
2006 V34 1: pp. 133–155
REAL ESTATE
ECONOMICS



A Simple Bayesian Procedure for Sample Size Determination in an Audit of Property Value Appraisals

Nathan Berg*

The article proposes a simple Bayesian technique for auditing property appraisals to determine whether state accuracy guidelines are met. The proposed technique addresses elicitation of appraisers' prior beliefs, computation of reappraisal sample sizes and reporting of audit results. To facilitate communication of quantitative audit findings to nonstatistician stakeholders, the concept of variance appears nowhere in prior elicitation or reporting. In contrast to classical frequentist techniques, the Bayesian procedure easily integrates expert judgment and responds flexibly to the arrival of new information. In addition, the Bayesian procedure significantly reduces the number of reappraisals required to regulate appraisal systems when they are functioning well. The technique can be applied in other settings where government officials audit their own work and must convince overseers, especially the public, that accuracy requirements are satisfied.

Public-sector property appraisers, whose responsibility is to compute annual property valuations for use by tax assessors and attest to their accuracy, are providers of technical expertise and what Walls and Quigley (2001) refer to as *socio-technical* services. Socio-technical services are those that require specialized communication skills for eliciting statistical information from nonstatisticians and for persuasively explaining quantitative issues to various stakeholders, in this case taxpayers, users of public services financed by property tax revenue and other constituencies in the local or regional political economy. The task of communicating algorithmic detail from the computation of property owners' tax bills and attesting to the effectiveness of quality control measures clearly fits the socio-technical label.

The focus of this article is the audit of public-sector property appraisals. Such audits require reappraisal of a relatively small number of properties using costly, in-depth appraisal methods and, thus, determination of an appropriate reappraisal sample size. The goal is to convince state officials and members of the

*School of Social Sciences, University of Texas at Dallas, Richardson, TX 75083-0688 and Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin 14195 Germany or nberg@utdallas.edu.

public that property appraisals satisfy exogenously given standards of accuracy such as those required by state law. The auditor, or appraisal authority, must therefore confront the difficult challenge of communicating to nonstatisticians about second-moment phenomena, namely, risk and dispersion.

Diverse voices in the scientific community have remarked on the challenges of risk communication and the benefits of simplicity in a variety of economically significant settings (Simon 1982, Slovic 2000, Gigerenzer 2002). Psychologists writing in *Science* (Hoffrage, Lindsey, Hertwig and Gigerenzer 2000) showed that logically equivalent descriptions of disease frequencies (*e.g.*, "three in 1,000" as opposed to 0.3%) in medical tests to screen for disease led patients to choose significantly different courses of treatment. In the design of learning systems in artificial intelligence, Simon, Valdes-Perez and Sleeman (1997) demonstrated that algorithmic complexity is often disadvantageous, not just because of computational costs, but because simple decision rules tend to be more robust in changing environments. Analyzing how economists construct persuasive arguments, McCloskey (1985) showed that the role of language reaches well beyond logical coherence and deductive chains relating axioms to theorems.

Confronting the communication issue in the context of an audit of property appraisals means having explanations for strategies used in determination of sample size, integration of expert judgment and the weighing of statistical benefits against rhetorical costs (in terms of algorithmic complexity) at the end-user stage. Complexity imposes costs whenever it impedes attainment of the audit's ultimate socio-technical goal, which is to quantitatively characterize appraisal accuracy in language that satisfies the constraints imposed by end-users' unfamiliarity with statistical jargon, including the concept of variance. In contrast to strictly technical property appraisal issues where few, if any, costs of complexity need be considered (*e.g.*, the inherent statistical challenges of estimating the market value of infrequently traded real assets with large and correlated location-specific components), tools designed for socio-technical tasks such as public-sector audits must deal explicitly with algorithmic complexity and its effects on end-users. Complexity not only increases skill requirements (possibly requiring direct expenditures on consultants or additional in-house personnel), it can also jeopardize the political legitimacy of quantitative decision-making procedures because of difficulty in justifying in-transparent black-box computations to nonexperts.

In the United States, United Kingdom and other European nations, legal definitions and customs concerning sufficiency of evidence are fundamentally ambiguous (Steele 1992). In practice, most auditors rely on rules of thumb, such as "choose $n = 30$," apparently with little justification. Some rely solely on expert judgment with virtually no statistical support.

In a number of U.S. states, state law specifies accuracy requirements for property appraisals. Given such exogenous requirements, compliance can be viewed as a binary outcome determined by comparing the allowable valuation error with the observed difference between two appraisals of the same property—one from an in-depth reappraisal regarded as a close approximation to true market value, the other derived from the standard, more economical appraisal procedure. In such cases, the main technical component of the audit problem is selecting an appropriate statistical model for the probability of noncompliance.

The Bayesian probability model proposed here has advantages in terms of both informational and cost efficiency. In contrast to audit methods based on classical statistics, it easily integrates the judgments of appraisal experts concerning local market conditions and information from previous audits. And because labor-intensive reappraisals are costly, the Bayesian procedure provides practical benefits by requiring smaller sample sizes in most cases.[1] Perhaps most important, the proposed procedure results in natural-language risk reporting derived from tail probabilities of the posterior distribution without invoking the concept of variance or other statistical jargon.[2]

The article is organized as follows. The next section reviews relevant work in the fields of property appraisal research, Bayesian audit methodology, Bayesian sample size determination, elicitation of expert judgment and risk communication. The third section describes classical approaches to sample size determination illustrating their limitations and, thus, the need for an alternative approach. The fourth section describes the article's main result, an algorithm for computing the minimum number of reappraisals required to achieve user-specified posterior confidence in the event of compliance. The fifth section presents examples and numerical results illustrating how the procedure works in practice. The final section concludes with a discussion of the broader issue of sufficiency of evidence in quality assurance tasks conducted on behalf of taxpayers and their representatives in local government.

[1] The claim depends on correct model specification.

[2] The audit problem analyzed in this article is based on a real-world compliance-reporting task for which the Dallas Country Appraisal District (DCAD) in Dallas, Texas, sought the author's advice. In 2001, DCAD faced the prospect of proving to the State Comptroller that Dallas County appraisals were within allowable limits for errors in property valuations. Thus, DCAD had to produce a politically persuasive and statistically grounded statement attesting to the accuracy of its appraisals. Among other issues, DCAD sought an answer to the sample size question: How many expensive in-depth reappraisals should be performed in order to satisfactorily check that the rate of compliance is close to 100%? Unsatisfied with answers provided by classical techniques, the Bayesian technique presented in this article was developed.

Background

There is compelling evidence that commercial and public-sector property appraisals are systematically biased (Geltner 1989, Graff and Young 1999, Shiller and Weiss 1999, Dietrich, Harris and Muller 2000). Some argue that gaps between statistical moments of appraisal-value and market-value distributions are rooted in psychological biases, such as anchoring effects (Clayton, Geltner and Hamilton 2001), although there is disagreement about their magnitude and economic significance (Diaz 1997). As a rule of thumb, appraisal dispersion as indicated by standard deviation appears to be approximately 10% (Hansz and Diaz 2003), although feedback, which enables learning, and experience (Spence and Thorson 1998) can moderate this dispersion somewhat. Appraisal bias is important not only for reasons relating to disputes over property tax collection, but also in eminent domain cases (Adams, Jackson and Cook 2001) and as a potential predictor of mortgage default (LaCour-Little and Malpezzi 2003). There are also important normative issues relating to appraisal returns and risk hedging, where overly smooth appraisal-based real estate time series can unfortunately mask the true covariance structure between real estate and other asset categories (Geltner 1989, Gau and Wang 1990, Hendershott and Kane 1995, Lai and Wang 1998, Gunnelin, Hendershott, Hoesli and Soderberg 2004).

The real estate literature on appraisal technique rests largely on several classic approaches (Isakson 1986, Lusht 1987, Kang and Reichert 1991, Isakson 1998, Pace 1998). According to Roulac, Adair, Crosby and Lim (2004), however, most contemporary appraisal research is difficult for real estate professionals to put into practice. One reason for the apparent gap between theory and practice seems to be insufficient appreciation of the simultaneous importance and difficulty of communicating about risk (O'Hagan 1998, Hoffrage, Lindsey, Hertwig and Gigerenzer 2000, Gigerenzer 2002). A similar gap between theory and practice (pointed out by Pham-Gia (1997) and O'Hagan (1998)) applies both to the Bayesian audit literature (Baker 1977, Menzefricke 1984, Rohrbach 1986, Tamura and Frost 1986, Laws and O'Hagan 2002) and sample size selection literature (Hora 1978, Aigner 1979, Cox and Snell 1979, Laws and O'Hagan 2000).

Sample size selection models with attractive features that resemble the model proposed in this article have been put forward (Chaloner and Duncan 1983, Pham-Gia and Turkkan 1992, Bernardo 1997, O'Hagan 1998, Wang and Gelfand 2002, Inoue, Berry and Parmigiani 2005). In general, this article's procedure differs from previous Bayesian sample size selection procedures in that it eschews the concept of variance in elicitation and reporting, it does not rely on normality assumptions or expected utility functions with special forms,

and it permits updating with small samples that lie on the boundary of sample space (*e.g.*, those with exactly zero noncompliant observations).

Steel (1992) points out that auditors are hired to provide an opinion. While courts have held auditors to certain standards of "reasonableness" and required a "rational basis" for their judgments, there is as yet no uniform standard of sufficient evidence in support of audit decisions. The proposed appraisal audit procedure aims to provide a new tool to better integrate quantitative analysis and judgment. It draws motivation from the unfortunate real-world prevalence of *ad hoc* sampling rules (*e.g.*, $n = 30$).

Classical Approaches

This section reviews classical or frequentist approaches to the sample size selection problem and, by way of contrast, demonstrates the advantages of Bayesian techniques in terms of informational, cost and socio-technical efficiency.

Hypothesis Tests

Define

$$Y_i = \begin{cases} 1 \text{ if the } i \text{ th unit audited is noncompliant} \\ 0 \text{ otherwise.} \end{cases}$$

$$Y_i = \begin{cases} 1 \text{ if the } i \text{ th unit audited is noncompliant} \\ 0 \text{ otherwise.} \end{cases}$$

Let p represent the probability that Y_i is noncompliant. All sampled units are assumed to be independent and have the same noncompliance probability p. **Q1** Define B as the number of noncompliant (*i.e.*, bad) units in a sample of n reappraised properties (so that $n - B$ is the number of compliant units). Under these assumptions, B has a binomial distribution with parameters p and n. According to standard Neyman–Pearson methodology, the null hypothesis H_0: $p = p_0$ is tested against the alternative hypothesis H_0: $p > p_0$ at the $(1 - \alpha)$ level by choosing a (minimal) critical value c such that

$$\Pr(B \leq c) = \sum_{i=0}^{c} \frac{n!}{i!(n-i)!} p_0^i (1 - p_0)^{n-i} \geq 1 - \alpha. \tag{1}$$

Because B is a count statistic, c is constrained to integer values. Noninteger values of c are required to exactly satisfy the condition Pr (No Type-I error)= $1 - \alpha$. Recognizing that the left-hand side of (1) is increasing in c for fixed n, it is obvious that the best critical value will be the minimal value of c satisfying the

inequality. Here α is the probability of incorrectly concluding that the population error rate is greater than p_0 under the null.

Without imposing further constraints, n and c are indeterminate because many combinations of c and n can be chosen to satisfy (1). For example, $(n,\ c) = (5,\ 0)$, $(35,\ 1)$ and $(82,\ 2)$ are all approximate 95% tests of the hypothesis $p = 0.01$.

A standard way to choose among an infinite list of 95% tests is to consider power against a particular simple alternative hypothesis. As in the continuous case, each critical-value/sample-size pair corresponds to a power function. Choosing the desired significance level and a point that is to intersect with the power function pins down n and c.

Accordingly, one requires that the power function satisfies:

$$\Pr(B > c) = 0.95 \quad \text{if } p = p_0 + 0.01. \tag{2}$$

In other words, when the true noncompliance rate is one percentage point higher than that specified by the null hypothesis, the test rejects the hypothesis p_0 95% of the time. Other values for the alternative noncompliance probability (perhaps higher than $p_0 + 0.01$) and the chance of rejection (perhaps lower than 95%) might also be reasonable and would reduce the required sample size.

The hypothesis test approach requires the user to provide the simple null p_0, the simple alternative p_1, the probability of type I error α as well as the power $1 - \beta$ at p_1. Imposing the power condition (2) simultaneously with the type-I error condition (1) jointly determines n and c. With $p_0 = 0.01$, $p_1 = 0.02$ and $\alpha = \beta = 0.05$, the solution to the system (1) and (2) is $n = 1567$ and $c = 22$. As mentioned above, the solution is only approximate, because n and c are constrained to be integers. At $p = 0.01$ and $n = 1567$, $\alpha = \Pr(B > 22) = 0.0478$. At $p = 0.02$, $\Pr(B \leq 22) = 0.0497$.

Thus, hypothesis testing leads to large reappraisal costs. By reducing the stringency of the confidence and power requirements to $\alpha = \beta = 10\%$, n and c are reduced correspondingly to $n = 945$ and $c = 13$. Different simple alternative hypotheses also may be substituted. For example, with $p_1 = 0.05$ in (2) rather than 0.02, the required sample size drops to $n = 181$ and $c = 4$.

In all cases, the number of reaudits is unappealingly large. No use is made of the appraiser's prior beliefs. And no probabilistic statement about the rate of noncompliance is possible because the procedure is based on classical methodology according to which the rate is not a random variable.

Confidence Intervals

Confidence interval estimation is another standard technique from classical statistics. The goal is to construct a random interval that covers the true parameter p 95% of the time. Because attention is focused on a rare event, p is near zero, and the one-sided confidence interval $[0,\ d(B,\ n)]$ is natural to consider, where $d(B,\ n)$ is the random upper endpoint of the interval defined implicitly by the coverage constraint: $\Pr(0 \leq p \leq d(B,\ n)) = 0.95$.

Unfortunately, explicit solutions for $d(B,\ n)$ do not exist, and indeterminacy once again requires additional constraints. A common solution to this problem is to constrain the length of the interval to a predetermined value. The selection of interval length is, however, difficult to link to beliefs, priors or expert judgment. What basis is there, for example, for choosing an interval length of 0.01 as opposed to 0.02?

Confidence Intervals Based on Normal Approximations

Confidence intervals based on transformations of the discrete variable B to asymptotic normality are also common. The most well known is:

$$\frac{B - np}{\sqrt{np(1-p)}} \overset{a}{\sim} \mathrm{N}(0, 1). \tag{3}$$

From this, it follows that the event $\{\frac{B}{n} - \frac{2}{n}\sqrt{np(1-p)} < p < \frac{B}{n} + \frac{2}{n}\sqrt{np(1-p)}\}$ occurs in approximately 95% of (large) samples drawn. Because the term $\sqrt{p(1-p)}$ attains a maximum of 0.5, the interval

$$\left[\frac{B}{n} - \sqrt{\frac{1}{n}},\ \frac{B}{n} + \sqrt{\frac{1}{n}}\right], \tag{4}$$

asymptotically covers p at least 95% of the time. By requiring (with little justification) an interval length of 0.02, the necessary sample size for an asymptotic 95% confidence interval is $n = 10{,}000$.

Required sample size can be reduced if one is willing to impose an upper bound $\phi \leq 0.5$ on p so that $\sqrt{p(1-p)}$ is bounded above by $\sqrt{\phi(1-\phi)}$. In general, if λ is the desired length of the confidence interval, ϕ is the upper bound of p, $1 - \alpha$ is the desired confidence level and $\Phi(.)$ is the standard normal cumulative distribution function, the sample size formula is:

$$n = 4\phi(1-\phi)(\Phi^{-1}(1-\alpha/2))/\lambda)^2. \tag{5}$$

Setting $\alpha = 0.05$, $\phi = 0.05$ and $\lambda = 0.02$ reduces the sample size to $n = 4(0.05)(0.95)(2/0.02)^2 = 1{,}900$. As impressive as this 81% reduction in sample

size is, 1,900 remains unacceptably large in the appraisal audit context, where as few as 100 reappraisals would vastly outstrip the resources available to most country appraisal offices in the United States.

A second commonly encountered normal approximation is

$$\arcsin\left(\sqrt{B/n}\right) \overset{a}{\sim} \mathrm{N}\left(\arcsin\left(\sqrt{p}\right), \frac{1}{4n}\right). \tag{6}$$

Rearranging the equation leads to the approximate 95% confidence interval:

$$\left[\sin^2\left(\arcsin\left(\sqrt{B/n}\right) - \sqrt{\frac{1}{n}}\right), \sin^2\left(\arcsin\left(\sqrt{B/n}\right) + \sqrt{\frac{1}{n}}\right)\right]. \tag{7}$$

Setting the length of this interval equal to 0.02 and solving for n (with numerical techniques) result in required sample sizes similar to those derived from the binomial model (*e.g.*, $n = 10{,}000$ with $\alpha = 0.05$).

The Bayesian Approach

The conjugate beta-binomial distributions are well known in Bayesian statistics (Chaloner and Duncan 1983, Calvin 1990, Wolfson and du Berger 1995). Suppose the noncompliance rate for the population of property appraisals is the random variable p with beta pdf

$$f_p(p) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} & \text{if p} \in (0, 1) \\ 0 & \text{otherwise,} \end{cases}$$

where a and b are parameters through which prior beliefs influence the distribution's shape. It is well known that $\mathrm{E}p = \frac{a}{a+b}$ and the mode of p is $(a-1)/(a+b-2)$ provided $a > 1$ and $b > 1$. Assuming that the appraisal compliance process at the level of individual sample units is *i.i.d.* with sample-unit noncompliance probability p, the number of noncompliant units B (the number of bad ones) in a sample of n reappraisals is binomial.[3]

The posterior pdf of p given realized values of B and n is

[3] The assumption that there is a common probability of noncompliance across all property appraisals may appear restrictive. Appraisers are likely to know of local conditions that could be used to hierarchically model the population by partitioning it into geographically defined subgroups that vary with respect to rates of compliance. Introducing additional complexity, however, undermines the goal of simple, transparent and politically robust risk communication. Although additional flexibility of the hierarchical approach across different stratification schemes would allow for the possibility of achieving even smaller sample sizes, the subsequent analysis shows that very small samples are satisfactorily achieved within the simpler framework.

$$f_{p|B,n} = \frac{\Gamma(n+a+b)}{\Gamma(B+a)\Gamma(n+b-B)} p^{B+a-1}(1-p)^{n+b-B-1}, \tag{8}$$

which has mean

$$\mathrm{E}[p \mid B, n] = \frac{a+B}{a+b+n} \tag{9}$$

and

$$\Pr(p \leq t \mid B, n) = \int_0^t \frac{\Gamma(n+a+b)}{\Gamma(B+a)\Gamma(n+b-B)} p^{B+a-1}(1-p)^{n+b-B-1} dp. \tag{10}$$

Additional user-supplied constraints are required to determine sample size. The next section considers how to formulate those constraints in a manner that efficiently elicits prior information from appraisal experts and leads to an intuitive statement about the accuracy of appraisals that is maximally user-friendly to nonstatisticians.

Prior Elicitation and Statement of Results

According to O'Hagan (1998), there is a surprising dearth of studies that strive to incorporate realistic priors.[4] O'Hagan expresses concern that Bayesian statisticians pay too little attention to elicitation and suggests that, when they do pay attention to elicitation, they should give more consideration to the importance of simple, familiar language when mapping stated beliefs (of those who possess prior information) into parameterizations of prior distributions. O'Hagan and colleagues rely on a specialized software package that conducts prior elicitation by asking experts (engineers in O'Hagan's case) to provide quantiles. Using visual feedback and a sequence of redundant questions, the software points out any inconsistencies to users among their responses, allowing for subsequent correction and continuing in a loop until the user chooses to terminate the elicitation process. The software then parameterizes the model based on the last prior distribution elicited or based on an average of implied parameterizations derived from multiple elicitations (Garthwaite and O'Hagan 2000).

User-supplied quantiles are not without problems, however. In a laboratory study, Hogarth (1975) found that interquartile ranges from elicited distributions contained considerably less than 50% of outcomes, implying that objective

[4] Bayesian researchers working on other aspects of the sample size selection problem also lament the lack of Bayesian field applications, similarly commenting on the high ratio of theoretical to applied work in the Bayesian paradigm (Pham-Gia 1997). Also see the opening quote of de Finetti in Hogarth (1975) imploring social scientists to develop improved techniques of prior elicitation.

distributions had higher dispersions than subjective priors derived from quartile elicitation. There is conflicting evidence about whether subject-area expertise reduces such bias (Stewart, Roebber and Bosart 1979, Christensen-Szalanski and Bushyhead 1981). O'Hagan (1998) conjectures that particular quantiles, for example tertiles (*i.e.*, partitioning the support into three equiprobable regions), are better suited for unbiased elicitation.[5]

Similar to O'Hagan (1998), the elicitation procedure here focuses on a particular quantile. The natural quantile to focus on is the lower probability tail of noncompliance $\Pr(p < p_0)$ cut off by the expert's prior point estimate p_0. Because prior p has a beta distribution, its shape is flexible and there seems to be little gain in averaging over multiple elicitations as in O'Hagan (1998), Walls and Quigley (2001, 2004) and Chaloner and Duncan (1983). The expert chooses both the quantile and its probability. The quantile is not abstract, as it concerns the rate of noncompliance which is the central motivation for the audit. Additional elicitation questions about upper tails or mid-distribution tertiles are far less intuitive and distant from the concerns of the expert in the context of the audit problem. While the one-shot elicitation advocated here may strike some as crude in comparison to elaborate elicitation techniques that appear elsewhere in the literature, procedural simplicity, the intuition of the expert and nonuniformity of expertise over the range of p argue in favor of a single question about the lower tail of the rate of noncompliance. Details of the elicitation procedure are provided next.

Prior Elicitation

The prior distribution is determined by asking the expert two questions:

> Q1: "In your opinion, what fraction of property appraisals are currently out of compliance? Please state your best guess." [Response denoted p_0.]
> Q2: "Realizing that the estimated rate of noncompliance you just mentioned (p_0) is uncertain and could actually be higher or lower, what is the chance, in your opinion, that the actual noncompliance rate is less than or equal to p_0?" [Response denoted k_0.]

Rather than asking the expert, who may be less than familiar with the statistical concept of variance, to state the precision of his or her prior beliefs in terms of

[5] Incentive problems may also lead to biased elicitation when state officials are in charge of auditing their own work. This article assumes, however, that a combination of positive incentives, reputation effects and reciprocal behavior (Gintis 2000) align the interests of appraisal officials with the goal of uncovering the objective frequency of noncompliance.

dispersion, Q2 asks for a confidence level or lower-tail probability k_0 measuring the (subjective) probability that the noncompliance rate is at least as small as p_0.

The elicitation literature raises the question of whether point estimates such as p_0 (elicited in Q1) should be equated with the mean or mode of the parametric distribution used to represent unknown p. O'Hagan (1998) argues that modes are more appropriate than means, especially for highly skewed distributions where the gap between the two can be large. Bernardo (1997) and Chaloner and Duncan (1983) also advocate elicitation using modes, although a number of classic articles in the Bayesian sample size literature, such as Pham-Gia and Turkkan (1992), Kadane, Dickey, Winkler, Smith and Peters (1980) and Bunn (1979), conduct elicitation by equating point estimates with parameterized means. Of course, the law of large numbers implies that the empirical noncompliance rate approaches the expected value of p as the sample size grows large. Because elicitation is stated in terms of this population characteristic rather than as a characteristic that applies to a single sample unit, it makes more sense to equate the point estimate elicited in Q1 with its expected value.

Thus, the expert's response to Q1 provides an estimate of the prior mean:

$$\frac{a}{a+b} = p_0. \tag{11}$$

The response from Q2 may be equated with the parameterized lower tail probability of the prior distribution:

$$\Pr(p \leq p_0; a, b) = \int_0^{p_0} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} dp = k_0. \tag{12}$$

Together, Equations (11) and (12) determine the gamma parameters a and b.[6] As an alternative, a multiple choice instrument covering the relevant ranges of p_0, k_0 and k_1 may be desirable.

The next step is to give symbolic expression to a natural English language statement about the posterior probability of compliance. Even with the prior

[6] Numerical grid search (coded in MATLAB and available from the author upon request) quickly computes values for a and b given user-supplied values of p_0 and k_0. The value of a that comes closest to satisfying $\Pr(p \leq p_0; a, b) = k_0$ while satisfying $\Pr(p \leq p_0; a, b) \leq k_0$ (so that the prior distribution parameters are no more confident than the stated level k_0) is selected. Note that $Pr(p \leq p_0; a, b)$ is not monotonic in b and it is therefore necessary to search over a large range. Another constraint built into the elicitation setup is the requirement that prior confidence level k_0 is at least 50%. If there is less than a 50% belief that $p \leq p_0$, then the expert should adjust p_0 upward so that there is better than a coin flip's chance that the true error rate is bounded above by the prior mean p_0.

distribution completely determined, an additional constraint on the posterior distribution is needed to solve for sample size. Because it is intuitive for non-statisticians and avoids the concept of variance, the following lower-probability-tail form for stating the statistical objective of the audit is proposed:

$$\Pr(p \leq 0.05 \mid \text{observed reappraisal sample}) = 0.90, \tag{13}$$

where the threshold 0.05 and target level of confidence 0.90 are chosen simply for illustration. This form provides an intuitive statement of the audit's main result: "I am 90% confident that the rate of noncompliance is under 5%." The remaining issue is what to assume about the observed reappraisal sample before it is observed. The discussion below makes clear that the remaining constraints that determine sample size correspond to varying strategies for dealing with uncertainty about the posterior lower tail probability.

Constraining the Posterior Distribution of p

In choosing a condition on the posterior distribution that targets the auditor's goal and provides scientific support for a quantitative statement summarizing audit results, the social dimension of the socio-technical task becomes critical. After all, if end-users do not believe, trust or understand the quantitative statements resulting from the appraisal audit, then the procedure has failed no matter how desirable its statistical properties.

Given the parameterized prior, the posterior distribution can be expressed as a function of the observed number of noncompliant units B and sample size n. Virtually all interesting statistics and probabilities computed from the posterior distribution will depend on B and n. Constraints must be imposed to determine n in advance of the observed value of B.

A variety of posterior distribution constraints have been suggested in the Bayesian sample size literature. One obvious approach is to turn to formal decision theory by specifying an expected utility function and choosing sample size as its optimizer with respect to n (Lindley 1997). However, the statistics literature has shown that special functional forms in the objective function usually fail to convince others as to their reasonableness compared with more straightforward interval or tail-probability constraints imposed directly on the posterior distribution (Adcock 1997, Joseph and Wolfson 1997, Pham-Gia 1997).

Among the alternatives to expected utility, one of the most commonly referred to is the average coverage interval, which selects sample size by minimizing n subject to the constraint that a fixed-length interval (with user-specified length) covers the posterior mean with average user-specified probability $1 - \alpha$.

When the posterior distribution is symmetric, this is a relatively straightforward calculation (Box and Tiao 1973), and Joseph, Wolfson and du Berger (1995) show how to handle the nonsymmetric case. A similar interval constraint that leads to different choices of sample size requires the *average* length of the $1 - \alpha$ coverage interval to equal user-specified l, while the coverage probability constraint holds exactly. The difference concerns whether l is fixed and the coverage probability constraint is imposed under the expectation operator with respect to possible samples (the first case), or whether the coverage probability is fixed and the coverage length constraint is imposed in expectation form (Joseph, Wolfson and du Berger, 1995).

An alternative to coverage length and coverage probability constraints in expectation is to require that desired posterior probabilities or interval lengths hold when the worst possible sample is observed—the so-called worst-outcome criterion (Pham-Gia and Turkkan 1992). As intuition might suggest, the worst-outcome criterion leads to more conservative (*i.e.*, larger) sample sizes. Bayes factors and power constraints on hypothesis tests have also been proposed in specialized applications with normality assumptions (Spiegelhalter and Freedman 1986, Weiss 1997). Despite the obvious limitations of normality assumptions, much of the existing sample size selection procedures rely on them (Joseph and Belisle 1997, Kadane and Wolfson 1998) or transforms to log normality to handle nonsymmetric distributions (Garthwaite and O'Hagan 2000).

While offering numerous possibilities, existing sample size determination techniques come with serious limitations for the purpose of the appraisal audit problem. Coverage intervals are ill suited for variables with highly skewed distributions and are confusing to explain to end-users. Given the asymmetry of p, normality assumptions are clearly inappropriate. Finally, existing beta-binomial techniques invoke a number of difficult-to-justify conditions that needlessly push the limits of the socio-technical constraints.

For example, Cox and Snell (1979) and Moors (1983) require users to provide two gamma distribution parameters without describing the mapping from simple language into those values. Bernardo's (1997) Bayesian sample selection technique adopts a beta-binomial approach and uses elicitation techniques similar to this article's; however, n is derived from expected utility maximization employing a difficult-to-justify information-theoretic expected utility objective. Chaloner and Duncan's (1983) sample selection procedure achieves admirable algorithmic simplicity (effectively satisfying the socio-technical constraints) in both elicitation of priors and specification of posterior constraints, but unfortunately relies on assumptions that rule out samples with exactly zero noncompliant units, a nonnegligible possibility in well functioning appraisal systems with small reappraisal samples. Pham-Gia and Turkkan's (1992) sample selection

procedure also shares similarities with this article's, including the beta-binomial setup. They warn, however, about problems when p is close to zero, because their coverage interval constraint depends on variance which, unlike the lower tail probabilities used in this article, is highly sensitive to changes in p close to zero.

In this article, the posterior constraint used to close the model and produce a minimal sample size derives from the following question:

Q3: "As you know, we are trying to decide how many reappraisals to conduct. The more the reappraisals, the more confident we can be about our statements concerning the rate of noncompliance. How confident would you like us to be about our statements? That is, what probability would you propose that our statements concerning noncompliance are correct?"

Having decided on a lower-tail-probability constraint (which can be viewed as a fixed length coverage interval $[0,\ t]$), the final question is whether the desired event $p \in [0,\ t]$ should hold with probability k_1 for the average sample, for all samples or for some sample in particular. When the audit system is functioning well, appraisal officials may expect to see exactly zero noncompliant units in n reappraisals (*i.e.*, $B = 0$). Building on this point of view that the natural default of the appraisal system is a well functioning state and that minimal checking is usually required, the procedure computes the minimum sample size required to achieve the condition

$$\Pr(p < t \mid B = 0, n) = k_1. \tag{14}$$

Solving Equation (14) in n yields the minimum possible sample size to achieve posterior confidence k_1 given the prior determined by responses to Q1 and Q2. Thus, the constraint used here to determine n is the opposite of Pham-Gia and Turkkan's (1992) worst outcome criterion.

One may question whether it is reasonable to select sample size based on the assumption that the best possible outcome will occur. Keep in mind, however, that the *ex post* statement the procedure delivers to those who ordered the audit reflects the posterior probability $\Pr(p < t \mid B)$, whether it achieves target posterior confidence k_1 (*i.e.*, if $B = 0$) or not. After the reappraisal sample is drawn and B is observed, the appraiser can make the simultaneously rigorous and intuitive claim that: The Appraisal Authority is $\Pr(p < t \mid B) \times 100\%$ confident that the rate of noncompliance is under the allowable limit t.

Happily, the Bayesian procedure allows the user to look at the posterior distribution after any number of reappraisals, make a statistically supported claim about Ep and $\Pr(p < t)$, reset priors at current posterior values and start

again. Starting again means using the procedure to compute a new minimum n conditional on the sample, or partial sample, observed to date. If $B > 0$ occurs, then, the new, more pessimistic prior is reset and the procedure is simply run again before drawing a new sample of reappraisals to test compliance.

The best case criterion reflects the outlook of Andrews and Smith (1983, p. 125) who described the advantages of Bayesian audit technique by noting that "small samples will suffice when the system is good but large samples are needed when it is bad." Indeed, this is precisely what is accomplished within the present framework, relying on expert intuition about whether the system is running well informed by prior p_0 relative to the exogenous threshold t. Furthermore, when the expert's intuition is not borne out by observation, the Bayesian technique recovers and updates. In contrast, the frequentist position leaves little room to accommodate intuition and adjust midstream according to observation.

Summary of the Appraisal Audit Procedure

The user provides:

- Prior expectation p_0 for the noncompliance rate, elicited as the response to Q1.
- Prior confidence k_0, the user's subjective lower tail probability that the true noncompliance rate p is below p_0, elicited as the response to Q2. If k_0 is 50% or less, whoever it is whose prior is being elicited should be encouraged to adjust p_0 and k_0 higher so that $k_0 > 0.50$. Note too that $k_0 < 1$ is required to avoid a degenerate prior.
- t, the threshold defining compliance, exogenously given by law, or provided by those requesting the appraisal audit. To guarantee 0% noncompliance requires every property in the district to be reappraised. To avoid this, it must be that $t > 0$. Note, too, that the audit is pointless if t is near 0.50 because, in that case, it guarantees nothing more than a coin toss's chance of compliance. Thus, the allowable threshold should be chosen considerably less than 50%: $t \ll 0.50$.
- desired level of posterior confidence k_1 associated with the event of population compliance, $p < t$, elicited as the response to Q3.

Next, the prior beta distribution of p is parameterized by solving for a and b in the two equations

$$p_0 = \frac{a}{a+b} \text{ and } \Pr(p < p_0) = k_0. \tag{15}$$

The minimum sample size n^* is computed by solving for n in:

$$is\Pr(p \geq t \mid B = 0; n) = k_1. \tag{16}$$

After n^* reappraisals have been conducted and rated as compliant or noncompliant, and if the realized number B is indeed zero, the audit concludes with the statement: The Appraisal Authority is $k1 \times 100\%$ confident that the rate of noncompliance is less than the allowable limit t.

For a strictly positive realized value of B, two options are available. First, by adjusting target confidence k_1 downward to the realized level of posterior confidence $\Pr(p < t \mid B =$ realized B), an analogous statement may be reported with "$\Pr(p < t \mid B =$ realized B) $\times$ 100% confident" replacing "$k_1 \times 100\%$ confident." This may be interpreted either as a moderate success, if $\Pr(p < t \mid B =$ realized B) is fairly close to k_1, or as a negative result if the observed reappraisals reveal that the population is highly unlikely to be in compliance.

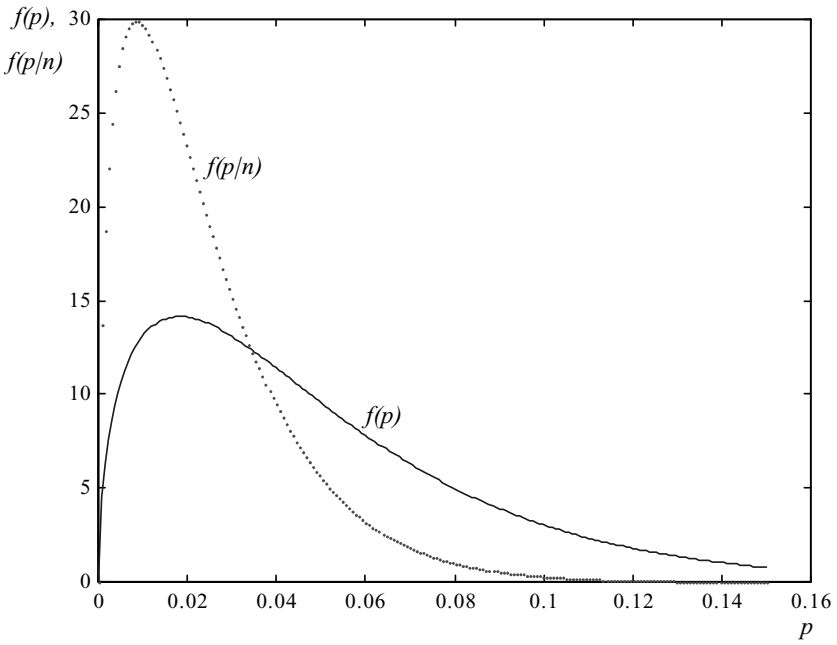
A second alternative available anytime a positive number of noncompliant reappraisals is observed is to halt the audit. At that point, the appraisal authority most likely modifies the standard appraisal technique and undertakes costly measures to move toward compliance. When desired, the prior p_0 may then be reset, either to the last posterior mean or by eliciting a new subjective prior. The procedure begins again and iterates until a satisfactory audit results.

An easy-to-use MATLAB code takes user-supplied inputs p_0, k_0, t and k_1 and provides the following output: prior distribution beta parameters a and b; minimal sample size n^* achieving the condition $\Pr(p < t \mid B = 0, n) = k_1$ and a user query to input the realized value of B so that the actual posterior probability of compliance $\Pr(p < t \mid B)$ can be computed. The next section presents several numerical examples illustrating the tremendous sample size savings possible using the procedure in conjunction with the best case criterion proposed above. The best case criterion is a formalization of Andrews and Smith's (1983) assertions about low sample auditing requirements for well-functioning systems. As argued above, the costs of being wrong about the well-functioning system are mitigated by the procedure's flexible response should noncompliant units in the initially small reappraisal sample be discovered.

Examples

Suppose the user's threshold for noncompliance is 5% ($t = 0.05$). Suppose also that the user believes the population is compliant ($p_0 = 0.05$), but only weakly so, reflected in the choice of prior probability $k_0 \equiv \Pr(p < 0.05) = 0.60$. The user wants to justifiably claim she is 90% confident ($k_1 = 0.90$) that the noncompliance rate is less than 5%. How many units must be audited?

Figure 1 ■ Prior and postaudit posterior pdfs, $f(p)$ and $f(p \mid n)$, for the rate of noncompliance p. The prior mean p_0 and exogenously given threshold t are set at 0.05. Prior weight on the event of compliance, k_0, is assumed to be 0.60. The target posterior weight on the event of compliance, conditional on $n = 33$ compliant observations, is $k_1 = 0.90$.



The prior beliefs p_0 and k_0 completely determine the prior beta distribution of p with parameter values $a = 1.52$ and $b = 28.88$. The posterior confidence requirement $\Pr(p < 0.05 \mid B = 0, n) = 0.90$ requires $n = 33$ reappraisals. If indeed all 33 are compliant, the appraiser may state, with rigor and simplicity, that, "I am 90% confident that the rate of compliance is above 95%." The MATLAB code quickly provides the mapping from beliefs into required sample sizes along with the graph depicted in Figure 1 containing an overlay of the prior (solid line) and posterior (dotted line) distributions of p (conditional on $B = 0$).

Obviously, the required sample size is sensitive to user-supplied inputs. Table 1 demonstrates this sensitivity. Its first row tabulates results of the numerical example described above that led to $n = 33$. Changing one user-supplied input at a time, Table 1 attempts to describe the sensitivity of n to p_0, k_0, t and k_1 in the neighborhood of that parameterization ($p_0 = t = 0.05$, $k_0 = 0.60$ and $k_1 = 0.90$). Setting the prior mean more pessimistically (p_0 from 5% to 10% noncompliant) only costs the auditor 10 additional reappraisals (required n

Table 1 ■ Required number of reappraisals (n) as a function of user-supplied inputs.

Required n	p_0	t	k_0	k_1
Increasing prior rate of noncompliance p_0				
33	0.05	0.05	0.60	0.90
37	0.06	0.05	0.60	0.90
40	0.07	0.05	0.60	0.90
41	0.08	0.05	0.60	0.90
42	0.09	0.05	0.60	0.90
43	0.10	0.05	0.60	0.90
Decreasing target noncompliance threshold t				
33	0.05	0.05	0.60	0.90
49	0.05	0.04	0.60	0.90
75	0.05	0.03	0.60	0.90
128	0.05	0.02	0.60	0.90
285	0.05	0.01	0.60	0.90
Increasing prior confidence k_0				
33	0.05	0.05	0.60	0.90
14	0.05	0.05	0.70	0.90
5	0.05	0.05	0.80	0.90
1	0.05	0.05	0.90	0.90
Decreasing prior confidence k_0				
33	0.05	0.05	0.60	0.90
37	0.05	0.05	0.59	0.90
41	0.05	0.05	0.58	0.90
47	0.05	0.05	0.57	0.90
55	0.05	0.05	0.56	0.90
65	0.05	0.05	0.55	0.90
Increasing posterior confidence k_1				
33	0.05	0.05	0.60	0.90
35	0.05	0.05	0.60	0.91
38	0.05	0.05	0.60	0.92
41	0.05	0.05	0.60	0.93
44	0.05	0.05	0.60	0.94
48	0.05	0.05	0.60	0.95

Note: The user-supplied inputs in the column headings are defined as follows. The prior rate of noncompliance is p_0; the exogenously given allowable limit for noncompliance is t; k_0 is the prior probability of the event $p < p_0$ and k_1 is the posterior probability that $p < t$.

ranging from 33 to 43). The second block of Table 1 demonstrates that changes in threshold t have a more dramatic effect. Aiming for $p < 0.01$ as opposed to $p < 0.05$ increases n from 33 to nearly 300. The third block of Table 1 shows, as one would expect, that increasing prior confidence k_0 reduces the number of required reappraisals. The last block of entries in Table 1 shows that increasing the required level of posterior confidence increases n, but not dramatically so: changing k_1 from 0.90 to 0.95 increases n from 33 to 48.

Apart from the question of sensitivity is the issue of what to do should noncompliant units appear in the reappraisal sample. If a noncompliant unit is observed at any time, the number of observed B out of n_0 re appraisals drawn so far can be fed back into the software as inputs, returning three quantities: the updated distribution of p (characterized by new values of a and b) with updated mean p_0, the updated value of $\Pr(p < p_0)$ and the new minimum sample size required to achieve $\Pr(p < t \mid B = 0, n) = k_1$.

For example, if after beginning with the same starting values as above, the 10th reappraisal is found to be noncompliant, the updated distribution of p has a larger mean, 0.0624. In this case, 80 additional reappraisals (as opposed to 33 before the first 10 reappraisals were observed) must be drawn in order to satisfy the posterior confidence condition.

Conclusion

This article argues that currently available audit methodology is not readily adaptable to the specific needs of government-appointed officials charged with auditing property appraisals and persuasively communicating the results to government overseers and the general public. The article develops a Bayesian procedure that is superior to classical hypothesis testing in terms of informational and economic efficiency. The proposed procedure systematically incorporates appraisal experts' prior beliefs. The procedure's final output enables auditors to issue a rigorous statistical statement about the rate of noncompliance and the degree of confidence ascribed to it. The procedure provides additional cost savings because its required sample sizes are small relative to those based on classical sample size determination formulas.

The article demonstrates how to compute the minimum number of reappraisals required to achieve a target level of confidence (*i.e.*, posterior probability) for the event that the rate of noncompliance is below the mandated threshold. If noncompliant reappraisals are observed, the procedure easily updates and recomputes the required number of reappraisals. Priors may be updated and sample sizes may be recomputed at any time because the procedure is Bayesian and therefore immune to classical problems associated with data mining and the distributional consequences of recomputing models multiple times in variant forms.

This article intentionally eschews the generality of elaborate parameterizations in favor of context specificity and simplicity. Given the socio-technical nature of the task, algorithmic complexity and statistical jargon tend to undermine the ultimate goal of persuading overseers and the public that appraisals comply with accuracy guidelines. The sample size determination procedure maps intuitively

articulated notions of accuracy and precision into the appropriate statistical concepts, and it maps quantitative audit results back into natural English language statements. Beyond the appraisal audit problem, it should be clear that the technique may be applied to similar situations in which government officials must audit their own work and convince overseers, especially the public, that exogenous standards of accuracy are met.

Thanks to James Murdoch for posing the reappraisal sample-size problem and helping in the development of its solution. Feedback from two anonymous referees is gratefully acknowledged. Valuable research assistance was provided by Yu Xue.

References

Adams, A.F., III, J.D. Jackson and J.P. Cook. 2001. Capital Market Theory and Real Estate Valuation: A Case Study in Choosing an 'Appropriate' Discount Rates. *Journal of Forensic Economics* 14: 119–133.

Adcock, C.J. 1997. Sample Size Determination: A Review. *The Statistician* 46: 261–283.

Aigner, D.J. 1979. Bayesian Analysis of Optimal Sample Size and a Best Decision Rule for Experiments in Direct Load Control. *Journal of Econometrics* 9: 209–221.

Andrews, R.W. and T.M.F. Smith. 1983. Pseudo-Bayesian and Bayesian Approach to Auditing. *The Statistician* 32: 124–126.

Baker, R.C. 1977. Determining Sample Size. *The Internal Auditor* 34: 36–42.

Bernardo, J.M. 1997. Statistical Inference as a Decision Problem: The Choice of Sample Size. *The Statistician* 46: 151–153.

Box, G.E.P. and G. Tiao. 1973. *Bayesian Inference in Statistical Analysis*. Addison-Wesley: Reading, MA.

Bunn, D.W. 1979. Estimation of Subjective Probability Distributions in Forecasting and Decision Making. *Technological Forecasting and Social Change* 14: 205–216.

Calvin, T.W. 1990. *ASQC Statistics Division How-To Series, Volume 7: How and When to Perform Bayesian Acceptance Sampling: Plus a Blast with TNT* Revised edition. ASQC Quality Press: Milwaukee, WI.

Chaloner, K.M. and G.T. Duncan. 1983. Assessment of a Beta Prior Distribution: PM Elicitation. *The Statistician* 32: 174–180.

Christensen-Szalanski, J.J.J. and J.B Bushyhead. 1981. Physicians' Use of Probabilistic Information in a Real Clinical Setting. *Journal of Experimental Psychology: Human Perception and Performance* 7: 928–935.

Clayton, J., D. Geltner and S.W Hamilton. 2001. Smoothing in Commercial Property Valuations: Evidence from Individual Appraisals. *Real Estate Economics* 29: 337–360.

Cox, D.R. and E.J Snell. 1979. On Sampling and the Estimation of Rare Errors. *Biometrika* 66: 125–132.

Diaz, J., III. 1997. An Investigation into the Impact of Previous Expert Value Estimate on Appraisal Judgment. *Journal of Real Estate Research* 13: 57–66.

Dietrich, J.R, M.S. Harris and K.A Muller III. 2000. The Reliability of Investment Property Fair Value Estimates. *Journal of Accounting and Economics* 30: 125–158.

Garthwaite, P.H. and A. O'Hagan. 2000. Quantifying Expert Opinion in the UK Water Industry: An Experimental Study. *The Statistician* 49: 455–477.

Gau, G.W. and K. Wang. 1990. A Further Examination of Appraisal Data and the Potential Bias in Real Estate Return Indexes. *Journal of the American Real Estate and Urban Economics Association* 18: 40–48.

Gelfand, A.E. and F. Wang. 2002. A Simulation Based Approach to Sample Size Determination under a Given Model and for Separating Models. *Statistical Science* 17: 193–208.

Geltner, D. 1989. Bias in Appraisal-Based Returns. *Journal of the American Real Estate and Urban Economics Association* 17: 338–352.

Gigerenzer, G. 2002. *Reckoning With Risk: Learning to Live With Uncertainty*. Penguin Books: London.

Gintis, H. 2000. *Game Theory Evolving*. Princeton UP: Princeton, NJ.

Graff, R.A. and M.S. Young. 1999. The Magnitude of Random Appraisal Error in Commercial Real Estate Valuation. *Journal of Real Estate Research* 17: 33–54.

Gunnelin, A., P.H. Hendershott, M. Hoesli and B. Soderberg. 2004. Determinants of Cross-Sectional Variation in Discount Rates, Growth Rates and Exit Cap Rates. *Real Estate Economics* 32: 217–237.

Hansz, J.A and J. Diaz, III. 2003. Valuation Bias in Commercial Appraisal: A Transaction Price Feedback Experiment. *Real Estate Economics* 29: 553–565.

Hendershott, P.H. and E.J. Kane. 1995. U.S. Office Market Values During the Past Decade: How Distorted Have Appraisals Been? *Real Estate Economics* 23: 101–116.

Hoffrage, U., S. Lindsey, R. Hertwig and G. Gigerenzer. 2000. Communicating Statistical Information. *Science* 290: 2261–2262.

Hogarth, R.M. 1975. Cognitive Processes and the Assessment of Subjective Probability Distributions. *Journal of the American Statistical Association* 70: 271–289.

Hora, S.C. 1978. Sample Size Determination in Bayesian Discriminant Analysis. *Journal of the American Statistical Association* 73: 569–572.

Inoue, L.Y.T., D.A. Berry and G. Parmigiani. 2005. Relationship between Bayesian and Frequentist Sample Size Determination. *American Statistician* 59: 79–87.

Isakson, H.R. 1986. The Nearest Neighbors Appraisal Technique: An Alternative to the Adjustment Grid Methods. *Journal of the American Real Estate and Urban Economics Association* 14: 274–286.

Isakson, H.R. 1998. The Review of Real Estate Appraisals Using Multiple Regression Analysis. *Journal of Real Estate Research* 15: 177–190.

Joseph, L., D.B. Wolfson and R. du Berger. 1995. Sample Size Calculations for Binomial Proportions via Highest Posterior Density Intervals. *The Statistician* 44: 143–154.

Joseph, L. and P. Belisle. 1997. Bayesian Sample Size Determination for Normal Means and Differences Between Normal Means. *The Statistician* 46: 209–226.

Joseph, L. and D.B. Wolfson. 1997. Interval-Based versus Decision Theoretic Criteria for the Choice of Sample Size. *The Statistician* 46: 145–149.

Kadane, J.B., J.M. Dickey, R.L. Winkler, W.S. Smith and S.C. Peters. 1980. Interactive Elicitation of Opinion for a Normal Linear Model. *Journal of the American Statistical Association* 75: 854–854.

Kadane, J.B. and L.J. Wolfson. 1998. Experiences in Elicitation. *The Statistician* 47: 3–19.

Kang, H. and A.K. Reichert. 1991. An Empirical Analysis of Hedonic Regression and Grid-Adjustment Techniques in Real Estate Appraisal. *Journal of the American Real Estate and Urban Economics Association* 19: 70–91.

LaCour-Little, M. and S. Malpezzi. 2003. Appraisal Quality and Residential Mortgage Default: Evidence from Alaska. *Journal of Real Estate Finance and Economics* 27: 211–233.

Lai, T. and K. Wang. 1998. Appraisal Smoothing: The Other Side of the Story. *Real Estate Economics* 26: 511–525.

Laws, D.J. and A. O'Hagan. 2000. Bayesian Inference for Rare Errors in Populations with Unequal Unit Sizes. *Applied Statistics* 49: 577–590.

——. 2002. A Hierarchical Bayes Model for Multilocation Auditing. *The Statistician* 51: 431–450.

Lindley, D.V. 1997. The Choice of Sample Size. *The Statistician* 46: 129–138.

Lusht, K.M. 1997. *Real Estate Valuation: Principles and Applications* Irwin: Chicago, IL.

McCloskey, D.N. 1985. *The Rhetoric of Economics*. University of Wisconsin Press: Madison, WI.

Menzefricke, U. 1984. Using Decision Theory for Planning Audit Sample Size with Dollar Unit Sampling. *Journal of Accounting Research* 22: 570–587.

Moors, J.J.A. 1983. Bayes' Estimation in Sampling for Auditing. *The Statistician* 32: 281–288.

O'Hagan, A. 1998. Eliciting Expert Beliefs in Substantial Practical Applications. *The Statistician* 47: 21–35.

Pace, R.K. 1998. Total Grid Estimation. *Journal of Real Estate Research* 15: 101–114.

Pham-Gia, T. 1997. On Bayesian Analysis, Bayesian Decision Theory and the Sample Size Problem. *The Statistician* 46: 139–144.

Pham-Gia, T. and N. Turkkan. 1992. Sample Size Determination in Bayesian Analysis. *The Statistician* 41: 389–392.

Rohrbach, K.J. 1986. Monetary Unit Acceptance Sampling. *Journal of Accounting Research* 24: 127–150.

Roulac, S., A. Adair, N. Crosby and L.C. Lim. 2004. The Emerging Global Real Estate Appraisal Research Agenda: Evidence from the ARES, ERES, PPRES and RICS Conferences. *Journal of Real Estate Literature* 12: 135–155.

Simon, H.A. 1982. *Models of Bounded Rationality*. MIT Press: Cambridge, MA.

Simon, H.A., R. Valdes-Perez and D.H. Sleeman. 1997. Scientific Discovery and Simplicity of Method. *Artificial Intelligence* 91: 177–181.

Shiller, R.J. and A.N. Weiss. 1999. Evaluating Real Estate Valuation Systems. *Journal of Real Estate Finance and Economics* 18: 147–161.

Slovic, P. 2000. *The Perception of Risk*. Earthscan: London.

Spence, M. and J.A. Thorson. 1998. The Effect of Expertise on the Quality of Appraisal Services. *Journal of Real Estate Research* 15: 205–215.

Spiegelhalter, D.J. and L.S. Freedman. 1986. A Predictive Approach to Selecting the Size of a Clinical Trial, Based on Subjective Clinical Opinion. *Statistics in Medicine* 5: 1–13.

Steele, A. 1992. *Audit Risk and Audit Evidence: The Bayesian Approach to Statistical Auditing*. Academic Press: London.

Stewart, T.R., P.J. Roebber and L.F. Bosart. 1997. The Importance of the Task in Analyzing Expert Judgment. *Organizational Behavioral and Human Decision Processes* 69: 205–219.

Tamura, H. and P.A. Frost. 1986. Tightening CAV (DUS) Bounds by Using a Parametric Model. *Journal of Accounting Research* 24: 364–371.

Walls, L. and J. Quigley. 2001. Building Prior Distributions to Support Bayesian Reliability Growth Modeling Using Expert Judgment. *Reliability Engineering and System Safety* 74: 117–128.

——. 2004. Structuring Engineering Knowledge to Inform Meaningful Prior Distributions for Modeling Reliability in Design—Principles and Practice. University of Strathclyde Working Paper 2004/10: Glasgow, UK.

Weiss, R. 1997. Bayesian Sample Size Calculations for Hypothesis Testing. *The Statistician* 46: 185–191.

Query

Q1 Author: Please check the repetition of the equation.