

Sample size selection for multiple samples-  
A brief introduction to credibility theory and  
an example featuring race-based insurance premiums

Stuart Klugman, Drake University  
For presentation at NYU – March 4, 2004

**Abstract**

For about one hundred years, actuaries have been concerned with the simultaneous estimation of parameters from numerous populations. Credibility procedures were designed to address this issue. Many years later these ideas were discovered by the statistics community. The first part of this paper reviews the credibility problem, provides a bit of history, and indicates the current popular solutions. The second part of this paper is an example where credibility results can be applied to a problem of simultaneously estimating 400 binomial proportion probabilities.

**Part I – Credibility history and motivation**

**1. The motivation for credibility**

Suppose you sell automobile insurance and the key quantity in setting premiums is the expected number of accidents per year for a given driver. Assume, for simplicity, that for a given driver the mean does not change over time and that the observed number of accidents per year is independent from year to year. Further assume that the number of accidents in a year for one driver is independent of that for another driver.

Five drivers have applied for insurance from your company. They are identical with regard to typical underwriting characteristics such as age, gender, type of car, location, etc. While you believe that the underwriting process will considerably reduce the variation from driver to driver, you do not believe that these five drivers have the same mean. To assist in the estimation process, you have four years of claims records for these drivers.

Number of claims in each year

Driver	Year			
	1	2	3	4
A	0	0	0	0
B	0	1	0	1
C	0	2	0	0
D	1	0	0	0
E	0	0	0	0

The task is to estimate the mean for each of the five drivers. Treating this as five separate problems and employing no prior information, would lead to the usual unbiased estimates of 0, 0.5, 0.5, 0.25, and 0. While there may be some solid statistical justification, these

answers cannot be used in practice. Drivers A and E would have premium of \$0, a result sure to keep their business and sure to lose money. Drivers B and C would face a considerable increase in premium and would be tempted to either find another insurer with better rates, or reduce their coverage. In either case, the losses on policies A and E would be not be recovered.

Alternatively, suppose there is considerable evidence from service bureaus that drivers in this underwriting category average 0.2 claims per year. We could claim that four years of data is not of much value and charge each driver the same premium. Drivers A and E would likely be able to find a company that offers a good driver discount, while drivers B and C will be delighted to remain our customer. Again, losses appear likely.

The answer, of course, is to set the premium somewhere in between. This is the essence of the credibility problem. The goal is to balance “luck” versus “skill”. For example, with regard to driver A, the company may prefer to offer a small reduction in premium, claiming the four years of no claims was mostly due to luck. However, driver A might argue for a large reduction, claiming the results are due to skillful avoidance of accidents.

## **2. Three standard settings for the use of credibility in insurance ratemaking**

The first setting was illustrated in the previous example. This is called individual risk rating and attempts to differentiate a single individual from other, similar, individuals.

The second is classification ratemaking. Once again, consider automobile ratemaking. The base premium may be based on gender (2 levels), age (3), type of car (5), and location (4). Thus there are 120 means to estimate. Some of the cells will have few observations and thus have unreliable sample means. In addition, there is likely to be some a priori structure to the results. For example, it might be desired to have a smaller estimate for those in the oldest age group for all levels of the other factors. It is more common in these problems to let the structure guide the solution rather than credibility theory. The generalized linear model has proved to be useful here.

The third is experience rating and is similar to the first setting. The difference here is that the mean to be estimated is for a group of people rather than a specific person. This was the problem that motivated the development of credibility theory in the 1900s. In workers' compensation insurance, premiums are based on the industry in which the employees work. A particular employer may expect a reduction in premium if their past history has been good (and will tolerate an increase if it has been bad). This is really no different from the first setting. However, it may be reasonable to assume that the individual employees who work for that employer are not homogeneous. The effect is to increase the variance of the observed results and thus makes the observations less reliable.

## **3. The two historical actuarial approaches**

By 1920, two methods had evolved for performing credibility analysis. Both are linear in

the following sense. Let  $x$  be an estimate of the mean based solely on the data from that person or group. It will typically be the sample mean. Let  $m$  be some a priori estimate of the mean (think of it as the value you would use if you had no data on that particular person). The credibility estimate is then  $Zx + (1 - Z)m$ , where  $Z$  is called the credibility factor and must be between 0 and 1.

The first approach has come to be called *limited fluctuation credibility* and is the most commonly used in practice. It is based on the same sample size techniques used in setting hypothesis tests and confidence intervals. It begins with the following question:

What is the smallest sample size that ensures  $x$  will be within  $r$  percent of the true mean  $p$  percent of the time?

For cases where the sample size exceeds the one that answers the question, set  $Z = 1$  and declare that the sample has full credibility. For cases where the sample size is less than that required, set  $Z$  equal to the square root of the ratio of the actual sample size to the required sample size.

There are two significant problems with limited fluctuation credibility. The first is that the values of  $r$  and  $p$  are arbitrary. This can make it difficult to justify the result to regulators or juries. The second is that the method ignores the other samples and also ignores the quality of  $m$ . It, too, is an estimated quantity and surely its accuracy is also relevant.

The standard for actuarial practice has been to assume a Poisson distribution and then use a normal approximation. For our original automobile example, the choice of the Poisson distribution is reasonable. With  $r = 5$  and  $p = 90$  it is not difficult to show that 1,082 expected claims are required. In our example, with about 0.2 claims per person expected, the required sample size for full credibility is 5,410. With 4 observations, the value of  $Z$  is 0.027 and past history will have almost no value.

The second approach has come to be called *greatest accuracy credibility*. The early proponents constructed a variety of arguments to support the equation  $Z = n/(n + k)$  where  $n$  is the sample size and  $k$  is an appropriately selected constant. This method recognizes that finite sample sizes are always inadequate. It was not until the 1967 that a rigorous derivation of this formula was constructed. In the context of our automobile drivers, suppose that within our underwriting class, the expected number of claims for a given driver is drawn from the distribution  $\pi(\theta)$ . Further assume (for simplicity) that a given driver's distribution of claim counts depends only on  $\theta$  (there may be other parameters, but they are the same for all drivers) with distribution  $f(x|\theta)$ . We then take a sample of size  $n$  from this driver,  $x_1, x_2, \dots, x_n$  and chose to restrict estimators to be of the form  $\hat{\theta} = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_n x_n$ . The coefficients are then selected to minimize  $E[(\hat{\theta} - \theta)^2]$  where the expectation is taken over everything that is random. The solution is

$$\hat{\theta} = Z\bar{x} + (1 - Z)\mu$$

$$Z = \frac{n}{n + v/a}$$

$$\mu = E(x) = E[E(x | \theta)]$$

$$v = E[Var(x | \theta)]$$

$$a = Var[E(x | \theta)].$$

There are a number of interesting features of this solution. First, because  $Var(x) = v + a$  we see that  $v$  and  $a$  apportion the variance into two pieces. The first,  $v$ , is called the expected process variance and represents variability due to “luck”. The second,  $a$ , is called the variance of the hypothetical means and represents variability due to skill. The credibility assigned to a given driver will be increased by taking a larger sample, decreasing the “luck” factor, or increasing the “skill” factor. It is important to note that none of these components are within the driver’s control. The two factors relate to the population of drivers. The insurance company can control the sample size and reduce the “skill” factor by creating a more homogeneous group.

Within the actuarial community, this particular result is sometimes called linear credibility (because of the restriction to linear estimators) and sometimes called Bayesian credibility (for no good reason). The latter name reflects the similarity to Bayesian estimation using squared error loss, but it needs to be emphasized that there is no prior distribution here, the “prior” distribution is real, just unobservable.

#### 4. A greatest accuracy credibility solution to the five driver problem

When implementing greatest accuracy credibility, the needed quantities can be obtained in a variety of ways, depending on the assumptions made and the type of information available.

Case 1 – All distributions are completely specified. Suppose, for example that given  $\theta$ ,  $x$  has the Poisson distribution with mean  $\theta$  and further suppose that from driver to driver  $\theta$  has a gamma distribution with parameters  $5/3$  and  $3/20$ . Then

$$\mu = E[E(x | \theta)] = E(\theta) = (5/3)(3/20) = 0.25$$

$$v = E[Var(x | \theta)] = E(\theta) = 0.25$$

$$a = Var[E(x | \theta)] = Var(\theta) = (5/3)(3/20)^2 = 0.0375$$

$$Z = \frac{4}{4 + 0.25/0.0375} = 0.375.$$

Drivers A and E would have an estimate of  $0.375(0) + 0.625(0.25) = 0.15625$ , which is a nice reduction, but not all the way to zero.

Case 2 – All distribution names specified but parameters estimated from the data. This will be used in the example to follow and will not be done here.

Case 3 – The process distribution is specified, but not the “prior” distribution. For the example, retain the Poisson distribution. Because  $\mu$  is  $E(x)$ , it can be estimated by the sample mean, 0.25. With the Poisson distribution,  $v = \mu$  and so can also be estimated as 0.25. With  $Var(x) = v + a$  and the sample variance being 0.2875, an estimate of  $a$  is 0.0375 and the same solution as in Case 1 results. It should be noted that I chose the parameters in Case 1 to produce this match.

Case 4 – No distributions are specified. A fully non-parametric approach can be constructed using some intuition to set the estimators and then adjusting them to be unbiased. For notation, let  $x_{ij}$  be the  $j$ th observation from the  $i$ th driver, let  $\bar{x}_i$  be the average for driver  $i$ , and let  $\bar{x}$  be the average over all drivers. Let  $r$  be the number of drivers and  $n$  be the number of observations for each driver. The estimators are

$$\hat{\mu} = \bar{x}$$

$$\hat{v} = \frac{1}{r(n-1)} \sum_{i=1}^r \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2$$

$$\hat{a} = \frac{1}{r-1} \sum_{i=1}^r (\bar{x}_i - \bar{x})^2 - \frac{\hat{v}}{n}.$$

It is possible for the estimate of  $a$  to be negative. In that case it is customary to set  $Z = 0$ . An interesting observation is that  $\hat{v}$  is the within mean square from an analysis of variance while the first term in  $\hat{a}$  is the between mean square divided by  $n$ . Thus negative values occur in the same situations when the  $F$  statistic is less than 1, which leads to acceptance of the null hypothesis. In this case, that means that all the drivers have the same mean. In such cases there is no point in applying a credibility factor because all drivers should have the same premium.

For the data in the example,  $\hat{\mu} = 0.25, \hat{v} = 19/60, \hat{a} = -1/60$  and so  $Z = 0$ .

## 5. Statistics references

In the 1970s shrinkage estimates become very popular in the literature. Several articles by B. Efron and C. Morris together<sup>1</sup> and separately explored shrinkage estimators. A favorite example was early season batting averages. After a month or two it is common to see a batting average over 0.400, but over 60 years have passed since it was last accomplished for a full season. As best as I can tell, interest in the general statistical community has waned. I can offer some possibilities:

1. The key to success is the tradeoff of bias for reduced mean square error. There are many cases where this may not be acceptable. For experience rating, we actually want biased estimates.
2. Simultaneous estimation may not be that common.
3. The availability of true Bayesian analysis of hierarchical models provides an alternative. For example, a true Bayesian solution to the driver example could be

---

<sup>1</sup> For an example of a jointly authored paper, see “Data Analysis Using Stein’s Estimator and its Generalizations,” *JASA*, 1975, vol. 70, pp. 311-319.

found using WinBugs with a Poisson-gamma model and vague priors on the gamma parameters.

## **Part II – An example with race-based pricing**

### **6. Introduction to the problem**

This example is based on a true story. The data have been fabricated as well as the identity of the minority group affected. Once upon a time, it was well-known that Elbonians had longer life-spans than non-Elbonians. As a result, it was more costly to provide annuities to Elbonians. However, prevailing social concerns prevented insurance companies from using Elbonian status as an underwriting factor. To work around this problem, some companies created two versions of the same policy, one with a high premium and one with a low premium. Agents were instructed to offer only the high premium policy to Elbonians and to offer only the low premium policy to non-Elbonians.

This activity took place some time ago and as time passed the offending companies went through many changes such as reorganization, changes in record keeping, and sales of the company itself or blocks of these policies. Today, Elbonians still living have brought a class-action suit against the offending companies and their successors. In theory, it should be easy to determine who should be party to the settlement – those who are Elbonian (assumed easy to establish) and who had purchased an annuity from an offending company (also easy to establish) at an inflated price (not so easy to establish). The problem is twofold:

1. Not all Elbonians received a high premium policy and not all non-Elbonians received a low premium policy.
2. There was nothing in the policy form itself that identified it as a high-premium policy designed for Elbonians.

However, for some reason, the agent did record on the application if the applicant was Elbonian or not.

It was thus decided that policy types would be divided into two categories. Those sold primarily to Elbonians (as indicated on the application) and those sold primarily to non-Elbonians. There would be no looking at the actual premium charged and no verification if the agent correctly identified the applicant. Thus, there are no data errors.

A particular company had issued 400 different policy types. The task is to identify which were sold primarily to Elbonians. This is to be done by sampling  $n$  policies from each type and recording the number with “Elbonian” written on the application. Those sold to  $r$  or more Elbonians would be declared policy types sold primarily to non-Elbonians.

## 7. An unsatisfactory solution

Let  $B$  define the cutoff that defines “primarily sold to Elbonians. That is, policies for which the true percent of applications from Elbonians is  $100B$  or more. Let  $p$  be the true proportion. Then the purpose of the sampling process is to be able to compare  $p$  to  $B$ . It is reasonable to set up the hypothesis test as

$$H_0 : p \geq B$$

$$H_1 : p < B.$$

That is because it is imperative that we make few errors in which no reparations are made when they should be. But the type II error of wrongly compensating those who don’t deserve it also need to be controlled. For example, suppose we set  $B$  at 0.90, the type I error probability at 0.05 and the type II error probability at 0.15 when  $B$  is 0.72. In that setting, a sample size of 37 is required and 30 or more Elbonians will be needed to declare the plan primarily for Elbonians.

There are several problems with this analysis. One is that the company had hoped for samples sizes of 10-15. A lot of money will be spent for 400 samples of size 37. Second, as with limited fluctuation credibility, the choices that led to this answer were arbitrary. Finally, the terms of the analysis are difficult to explain. They relate to errors of the procedure. They do not give any indication of the ultimate accuracy of the process.

## 8. A credibility based approach

To help refine the analysis, samples of size 5 were taken from each of the 400 populations. The number of Elbonians in each is given below.

Number of Elbonians	Number of plans
0	155
1	27
2	17
3	26
4	28
5	147

It is clear that, for the most part, policies either had a very high or very low proportion of Elbonians.

The analysis below uses a fully parametric model and maximum likelihood estimation of the parameters. The model for the distribution of the true value of  $p$  is the beta distribution:

$$\pi(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}, 0 < p < 1.$$

This assumption is both for convenience and because its shape seems reasonable. For the observed data, the conditional distribution is binomial:

$$f(x|p) = \binom{5}{x} p^x (1-p)^{5-x}, x = 0, \dots, 5.$$

The marginal distribution is

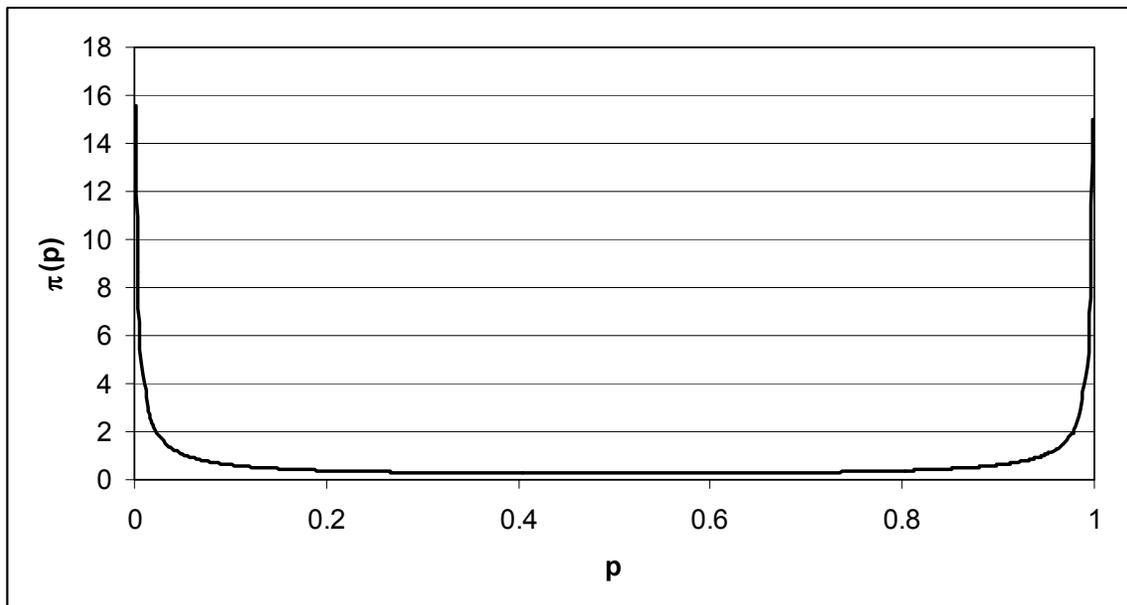
$$\begin{aligned} f(x) &= \int_0^1 \binom{5}{x} p^x (1-p)^{5-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1} dp \\ &= \frac{5! \Gamma(a+b) \Gamma(x+a) \Gamma(5-x+b)}{x! (5-x)! \Gamma(a) \Gamma(b) \Gamma(5+a+b)}, x = 0, \dots, 5. \end{aligned}$$

The maximum likelihood estimates are

$$\hat{a} = 0.153113$$

$$\hat{b} = 0.158898.$$

A chi-square goodness-of-fit test yields a test statistic of 2.11 with a  $p$ -value of 0.55, indicating strong support for this model. A graph of this distribution appears below.



It confirms supports our opinion that the majority of plans are at one extreme or the other.

The usual credibility approach would be to use this model to provide improved estimates. Instead, I have elected to focus on error rates. A reasonable goal would be to minimize the expected number of errors made when classifying 400 policy types. For a given type, the probability of erroneously classifying a policy as intended for non-Elbonians is

$$\begin{aligned}
\Pr(x < r, p \geq B) &= \sum_{i=0}^{r-1} \Pr(x = i, p \geq B) \\
&= \sum_{i=0}^{r-1} \int_B^1 \binom{n}{i} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{i+a-1} (1-p)^{n-i+b-1} dp \\
&= \sum_{i=0}^{r-1} \binom{n}{i} \frac{\Gamma(a+b)\Gamma(i+a)\Gamma(n-i+b)}{\Gamma(a)\Gamma(b)\Gamma(n+a+b)} [1 - \beta(B; i+a, n-i+b)]
\end{aligned}$$

where  $\beta(B; i+a, n-i+b)$  is the incomplete beta function. This is not a significance level, because it is not conditioned on the null hypothesis being true. The other error probability is

$$\Pr(x \geq r, p < B) = \sum_{i=r}^n \binom{n}{i} \frac{\Gamma(a+b)\Gamma(i+a)\Gamma(n-i+b)}{\Gamma(a)\Gamma(b)\Gamma(n+a+b)} \beta(B; i+a, n-i+b).$$

After checking out a variety of possibilities, the following emerged as a good choice:  $B = 0.8$ ,  $n = 10$ , and  $r = 8$ . This will cause 1% of Elbonian plans to be classified as non-Elbonian and 3% of non-Elbonian plans to be classified as Elbonian.

## 9. Final comments

This solution met all the key requirements of the problem. The sample size is modest and the performance of the method can be explained in terms of quantities that are important to the affected parties.

Equal sample sizes are not required and it would even be possible to construct an updating mechanism. That is, maybe if 5 out of 5 are Elbonian, stop sampling and declare the policy type to be primarily for Elbonians. If it turns out otherwise, sample 5 more, and so on until the chance of error is nearly eliminated. The conjugate nature of the beta and binomial distributions makes this relatively straightforward.

It may be that in the populations there are different (known) numbers of plans of each type. A better criterion might be to control the minimum number of policies expected to be classified in error. This would require the sample size to vary with the population size.