ACTUARIAL RESEARCH CLEARING HOUSE
1994 VOL. 1

OPERATIONAL BOOTSTRAPPING FOR THE LEAST SQUARES CREDIBILITY Z

William A. Bailey

Kemper National Insurance Companies

Long Grove, Illinois 60049-0001

ABSTRACT OF THE ABSTRACT

In Mr. Gary G. Venter's paper on Credibility in Foundations Of Casualty Actuarial Science he suggeted that "estimating the variance of the estimated [least squares credibility] Z would help provide an understanding of the accuracy of the calculation..." A usual estimator for the complement of the least squares credibility is $\frac{S}{n \cdot \bar{T}}$.

My paper studies in some detail his Example 5.2 (pp. 433-434), which involves 9 risks each subject to its own Gamma distribution:

(1) Bivariate and trivariate numerical generalized convolutions are used to obtain distributions of $\frac{S}{n \cdot \bar{T}}$ for n=6,12 and 18, given the distributions to which each risk is subject;

(2) Operational bootstrapping is used to generate a bootstrap distribution of $\frac{S}{n \cdot \bar{T}}$ for n=18, given Mr. Venter's 18 years of random-number-generated experience for each of the 9 risks;

(3) A special algorithm is described (sketched) for using bivariate numerical generalized convolutions to sample and resample from a Gamma distribution, the sampling and resampling is handled implicitly and without random numbers. An $\frac{n-1}{n}$ bias-correction factor is suggested (and confirmed by Mr. Charles S. Fuhrer, FSA, of Washinton National) for use where the bootstrapping is for variance.

KEYWORDS: Least squares credibility, operational bootstrapping,

bivariate and trivariate numerical generalized convolutions, sampling and resampling without random numbers, efficiency of the bootstrap.

ABSTRACT

We look at least squares credibility from three viewpoints:

(1) **Distributions of The Least Squares Credibility Given A Distribution For Each Risk.** Given that each of N risks is subject each year for n years to its own known (discretized) distribution use generalized bivariate numerical convolutions to find the bivariate distribution of $(X_{g\cdot} , S_{g\cdot})$ for the g^{th} risk $(g=1,2,\ldots,N)$; use generalized trivariate numerical convolutions to find the trivariate distribution of

$$(S_{\cdot\cdot} , X_{\cdot\cdot} , T_{\cdot\cdot}) = (\sum_{g=1}^{N} S_{g\cdot}/N , X_{\cdot\cdot} , \sum_{g=1}^{N} (X_{g\cdot}-X_{\cdot\cdot})^2/(N-1)),$$

restrict attention to the bivariate marginal distribution of $(S_{\cdot\cdot} , T_{\cdot\cdot})$, and transform this bivariate distribution into the univariate distribution of $\frac{S_{\cdot\cdot}}{n\cdot T_{\cdot\cdot}}$; and calculate the expected values $E[\frac{S_{\cdot\cdot}}{n\cdot T_{\cdot\cdot}}]$. It may be noteworthy that, since $S_{\cdot\cdot}$ and $T_{\cdot\cdot}$ are not necessarily independent random variables, $E[\frac{S_{\cdot\cdot}}{n\cdot T_{\cdot\cdot}}]$ is not necessarily equal to $\frac{E[S_{\cdot\cdot}]}{n\cdot E[T_{\cdot\cdot}]}$.

(2) **Operational Bootstrapping For Least Squares Credibility.** Given that N risks (each subject to its own unknown distributions) have been observed each year for n years use generalized numerical convolutions to generate for each

g (g=1,2,...,N), the bivariate distribution of ($X^*_{g.}$, $S^*_{g.}$), i.e. resample means and resample variances from resamples of size n taken from the observed samples (experience) of size n, form the distribution of ($S^*_{i.}$, $X^*_{i.}$, 0), use generalized numerical convolutions to generate the trivariate distribution of

$$(S^*_{..}, U^*, T_{..}) = (\sum_{g=1}^{N} S^*_{g.}/N , X^*_{..} , \sum_{g=1}^{N} (X^*_{g.}-X^*_{..})^2/(N-1)),$$

focus on the marginal bivariate distribution of

$$(S^*_{..} , T^*_{..})$$

and transform this distribution into the univariate distribution of

$$\frac{S^*_{..}}{n \cdot T^*_{..}},$$

from which we can calculate the expected value $E[\frac{S^*_{..}}{n \cdot T^*_{..}}]$. This procedure is referred to as operational bootstrapping for least squares credibility.

Comparing $E[\frac{S^*_{..}}{n\, T^*_{..}}]$ from (2) with $\frac{s^2}{s^2+n \cdot t^2}$ shows how well the operational bootstrapping for least squares worked with the given observed data to estimate the complement (1-Z) of the least squares credibility.

(3) **Efficiency of operational bootstrapping for the least squares credibility.** Given that each of N risks is subject each year for n years to its own known distribution we use generalized numerical convolutions to generate for each g (g=1,2,...,N) the distribution of

$$(X^{**}_{g\cdot} , S^{**}_{g\cdot})$$

where each of these pairs consists of a mean and variance of a possible resample of size n taken from a possible sample of size n taken from the known distribution for Risk g. This step requires a rather complicated algorithm to handle (implicity and without random numbers) the required sampling and resampling. The rest of the procedure is identical to (2) above, having replaced

$$(X^{*}_{g\cdot} , S^{*}_{g\cdot}) \text{ with } (X^{**}_{g\cdot} , S^{**}_{g\cdot}).$$

The statistical efficiency of operational bootstrapping for the least squares credibility can be determined by comparing

$$E[\frac{S^{**}_{\cdot\cdot}}{n\cdot T^{**}_{\cdot\cdot}}] \text{ from (3) with } \frac{s^2}{s^2+n\cdot t^2}.$$

Table Of Contents

Section 1. BACKGROUND MATERIAL

Our basic source material comes from the section on Least Squares Credibility in the chapter on Credibility in the book The Foundations of Casualty Actuarial Science, which was published by the Casualty Actuarial Society in 1990. My sincere appreciation to Gary Venter and the Casualty Actuarial Society for permission to reproduce (with minor changes) some of his material in this section of my paper.

It is supposed that the loss for N risks are observed for a period of n years. The pure premium (losses divided by exposures) of the g^{th} risk in year u is denoted as X_{gu}. The g^{th} risk is assumed to be subject to some probability distribution f_{X_g}. The pure premium, say C_{g0}, for a future time period (referenced as time 0) is to be estimated for the g^{th} risk, using the formula $C_{g0} = Z \cdot X_{g\cdot} + (1-Z) \cdot X_{\cdot\cdot}$; where $X_{g\cdot}$ is the average observed pure premium for risk g over the n years, $X_{\cdot\cdot}$ is the grand average of all the risks for those n years, and Z is a least squares credibility factor chosen so as to minimize

$$E[X_{g0} - (Z \cdot X_{g\cdot} + (1-Z) \cdot X_{\cdot\cdot})]$$

In formulas,

$$X_{g\cdot} = \sum_u X_{gu}/n, \text{ and}$$

$$X_{\cdot\cdot} = \sum_g X_{q\cdot}/N.$$

It turns out that

$$Z = \frac{n}{(n + \frac{s^2}{t^2})},$$

$$\text{so } 1-Z = \frac{s^2}{(s^2+n\cdot t^2)},$$

where s^2 is the average variance of an individual risk over time, and t^2 is the variance across the risks of the individual risk means.

One approach is to estimate s^2 and t^2 by first calculating the statistics

$$S_{g\cdot} = \sum_{u} (X_{gu}-X_{g\cdot})^2/(n-1)$$

$$S_{\cdot\cdot} = \sum_{g} S_{g\cdot} /N \quad \text{and}$$

$$T_{\cdot\cdot} = \sum_{g} (X_{g\cdot}-X_{\cdot\cdot})^2/(N-1).$$

Then, since

$S_{\cdot\cdot}$ is an unbiased estimator of s^2,

$n\cdot T_{\cdot\cdot}$ is an unbiased estimator of $s^2+n\cdot t^2$,

$T_{\cdot\cdot}-S_{\cdot\cdot}/n$ is an unbiased estimator of t^2, and

$$1-Z = \frac{s^2}{(s^2+n\cdot t^2)},$$

we could use $\frac{S_{\cdot\cdot}}{n\cdot T_{\cdot\cdot}}$ as an estimator for

$$1-Z = \frac{s^2}{(s^2+n\cdot t^2)}.$$

I have embellished Mr. Venter's symbols S_q, S and T to be $S_{q..}$, $S_{..}$ and $T_{..}$, respectively.

Section 2. INTRODUCTION

One objective of this paper is to study a particular hypothetical portfolio of risks to determine how well $\frac{S_{..}}{n \cdot T_{..}}$ can be expected to do as an estimator of 1-Z. The hypothetical portfolio of risks will be that described in Example 5.2 on pages 433-434 of Venter (1990); namely, nine risks each subject to its own Gamma distribution:

Risk	b	r	b·r	$b^2 \cdot r$
1	.6159	1.0476	.6452	.3974
2	.8001	0.9063	.7251	.5802
3	.6098	0.9654	.5887	.3590
4	.2391	0.9219	.2204	.0527
5	.5206	1.0184	.5302	.2760
6	.6768	1.0937	.7402	.5010
7	.9575	1.1395	1.0911	1.0447
8	.1999	1.0153	.2030	.0406
9	.5083	9.9320	.4737	.2408
			.5797	.3880

From this table we can calculate s^2=.3880 and t^2=.0747 (dividing by N-1=8), and find that

n	$\frac{s^2}{s^2+n \cdot t^2}$
6	.464
12	.302
18	.224

Although I could have used approximate integration to obtain a representation of the Gamma distribution for each of the 9 risks, I used the package @RISK

from Palisades to generate 4096 outcomes from each of the 9 Gamma distributions. In order to produce correct means and variances for the representations of each of the Gamma distributions, I adjusted the outcomes by

(a) multiplying by the square root of the ratio of the population variance to the variance of the sample, and

(b) adding the excess of the population mean over the mean of the adjusted sample.

Replacing any resulting negative amounts with zero did not materially affect the resulting distributions.

Section 3. DISTRIBUTION OF LEAST SQUARES CREDIBILITY

In this section we are assuming that we know the differing distributions to which each of N risks are subject, and we are not using random-number-generated experience like that generated in Venter (1990).

A method for performing generalizec bivariate numerical convolutions is described in ref [2] and is available in the computer package COCONUT™ from MathWare. The procedure using generalized numerical convolutions to generate the distribution of the least squares credibilities is as follows:

Let f_{X_g} denote what is, in effect, the discretized representation of the distribution for Risk g. Transform each of these univariate distributions into the corresponding bivariate distribution f_{X_g, X_g^2} replacing each x_g in f_{X_g} by the number pair (x_g , x_g^2). Convolute together n bivariate distributions

$$f_{X_{g1}} , f_{X_{g2}} , \ldots , f_{X_{gn}}$$

where each $f_{X_{gu}}$ (u=1,2,...,n) is independently and identically distributed as f_{X_g}, to obtain the bivariate distribution of

$$(X_{g\cdot} , S_{g\cdot}) = (X_{g\cdot} , \sum_{u=1}^{n} (X_{gu}-X_{g\cdot})^2/(n-1)).$$

This can be accomplished by performing n-1 generalized convolutions, to obtain recursively the bivariate distributions of

$$\left(\sum_{u=1}^{1} X_{gu}/1 \ , \ 0 \right)$$

$$\left(\sum_{u=1}^{2} X_{gu}/2 \ , \ \sum_{u=1}^{2} \left(X_{gu} - \sum_{u=1}^{2} X_{gu}/2\right)^2/(2-1) \right)$$

$$\left(\sum_{u=1}^{3} X_{gu}/3 \ , \ \sum_{u=1}^{3} \left(X_{gu} - \sum_{u=1}^{3} X_{gu}/3\right)^2/(3-1) \right)$$

.

.

.

$$\left(\sum_{u=1}^{n} X_{gu}/n \ , \ \sum_{u=1}^{n} \left(X_{gu} - \sum_{u=1}^{n} X_{gu}/n\right)^2/(n-1) \right) = \left(X_{g\cdot} \ , \ S_{g\cdot} \right)$$

Thus we obtain the distribution of $(X_{g\cdot} , S_{g\cdot})$ for each $g=1,2,\ldots,N$.

For each $g=1,2,\ldots,N$ we transform the distribution of $(X_{g\cdot} , S_{g\cdot})$ into the distribution of

$$(S_{g\cdot}, X_{g\cdot} , 0).$$

Next we obtain the distribution of

$$\left(\sum_{g=1}^{N} S_{g\cdot}/N \;,\; \sum_{g=1}^{N} X_{g\cdot}/N \;,\; \sum_{g=1}^{N} (X_{g\cdot} - \sum_{g=1}^{N} X_{g\cdot}/N)^2/(N-1) \right)$$

by transforming the bivariate distribution of $(X_{1\cdot}, S_{1\cdot})$ into

$$(S_{1\cdot}, X_{1\cdot}, 0),$$

recursively performing N-1 trivariate generalized numerical convolutions to obtain the trivariate distributions of

$$\left(\sum_{g=1}^{2} S_{g\cdot}/2 \;,\; \sum_{g=1}^{2} X_{g\cdot}/2 \;,\; \sum_{g=1}^{2} (X_{g\cdot} - \sum_{g=1}^{2} X_{g\cdot}/2)^2/(2-1) \right)$$

.

.

.

$$\left(\sum_{g=1}^{N} S_{g\cdot}/N \;,\; \sum_{g=1}^{N} X_{g\cdot}/N \;,\; \sum_{g=1}^{N} (X_{g\cdot} - \sum_{g=1}^{N} X_{g\cdot}/N)^2/(N-1) \right) = (S_{\cdot\cdot}, X_{\cdot\cdot}, T_{\cdot\cdot})$$

transforming this trivariate distribution into the bivariate distribution of

$$(S_{\cdot\cdot}, T_{\cdot\cdot})$$

and this distribution into the univariate distribution of

$$\frac{S_{\cdot\cdot}}{n \cdot T_{\cdot\cdot}}.$$

$X_{g\cdot}$ and $S_{g\cdot}$ are not in general independent, so we cannot expect $S_{\cdot\cdot}$ and $T_{\cdot\cdot}$ to be independent; thus, $E[\frac{S_{\cdot\cdot}}{n \cdot T_{\cdot\cdot}}]$ is not in general equal to $\frac{E[S_{\cdot\cdot}]}{n \cdot E[T_{\cdot\cdot}]}$.

The algorithm for the generalized trivariate numerical convolutions is a natural extension of the algorithm described in Bailey (1993) for generalized bivariate numerical convolutions. It is available in the computer package COCONUT™ from MathWare.

$$\frac{S_{..}}{n \cdot T_{..}}$$

	cum	n=6	n=12	n=18
	.000001	.05	.05	.05
	.00001	.06	.07	.06
	.0001	.09	.08	.07
	.001	.11	.10	.09
	.01	.18	.13	.11
	.025	.19	.15	.12
	.05	.23	.17	.13
	.1	.25	.20	.15
	.2	.31	.23	.18
	.3	.37	.25	.19
	.4	.44	.28	.21
	.5	.50	.31	.23
	.6	.55	.35	.24
	.7	.66	.39	.28
	.8	.73	.45	.31
	.9	.96	.53	.36
	.95	1.09	.57	.41
	.975	1.34	.62	.45
	.99	1.34	.73	.47
	.995	1.49	.77	.48
	.9999	2.16	1.04	.68
	.99999	2.20	1.09	.75
	.999999	3.26	1.12	.82
(1)	E[S/(n·T)]	.555	.340	.246
	Var[S/(n·T)]	.079	.017	.007
(2)	E[S]/n·E[T]	.464	.302	.224
	(1)/(2)	1.207	1.126	1.098

The last row in this table gives the degree of bias inherent in using $\frac{S_{..}}{n \cdot T_{..}}$ as an estimator for 1-Z, at least for the hypothesized portfolio. The row identified with a (2) is also equal to $\frac{s^2}{s^2+n \cdot t^2}$.

Section 4. OPERATIONAL BOOTSTRAPPING FOR LEAST SQUARES CREDIBILITY

The procedure using generalized numerical convolutions to generate the bootstrap distribution of the least squares credibilities is similar to that described above in the Section Distributions of Least Squares Credibility, except that for each g=1,2,...,N we let f_{X_g} be the empirical distribution of losses implied by random-number-generated experience of Risk g with n=18.

Portions of the following random-number-generated "experience" not available on page 433 in Venter (1990) were obtained directly from Gary Venter.

3 Random-Number-Generated Trials of 6 Years

Combined Into One Trial of 18 Years

Risk

Year	#1	#2	#3	#4	#5	#6	#7	#8	#9
1	.430	.247	.661	.182	.311	.301	.219	.002	.796
2	.375	1.587	.237	.351	.664	.253	1.186	.058	.260
3	2.341	1.939	063	.011	1.002	.044	.431	.235	.932
4	.175	.712	250	.022	.038	.109	1.405	.018	.857
5	1.016	.054	.602	.019	.370	2.105	.241	.713	.129
6	.466	.261	.700	.252	2.502	.891	.804	.208	.349
7	.215	.643	.121	.156	.141	1.255	1.993	.182	.089
8	.111	.405	.723	.008	.486	1.900	.286	.077	.484
9	.287	.234	.660	.410	.031	.517	2.261	.226	.432
10	.824	1.188	1.493	.226	1.109	.243	.897	.001	.546
11	1.340	2.030	.220	.292	1.002	1.229	.955	.600	.185
12	.237	.726	.066	.555	.685	1.215	.427	.286	1.413
13	.884	.190	.276	.169	.587	.819	1.515	.145	.254
14	.878	.135	.051	.153	.245	.055	.230	.111	.341
15	.159	.010	.539	.024	2.150	.049	2.099	.111	.296
16	.054	.092	.554	.075	.293	.313	.242	.165	.155
17	.470	.619	.039	.209	.478	.908	2.261	.476	.031
18	.147	.027	.152	.322	.359	.872	2.384	.188	.214
$X_{i\cdot}$	.578	.617	.412	.191	.692	.727	1.102	.211	.431
S_i	.309	.401	.129	.023	.431	.375	.618	.037	.122

$$T_{\cdot\cdot} = .071$$

$$S_{\cdot\cdot} = .272$$

$$\frac{S_{\cdot\cdot}}{n \cdot T_{\cdot\cdot}} = .213 \text{ where } n=18$$

$$\frac{S^{*}_{..}}{n \cdot T^{*}_{..}}$$

cum	n=18
.000001	.03
.00001	.04
.0001	.05
.001	.06
.01	.08
.025	.09
.05	.10
.1	.11
.2	.13
.3	.14
.4	.15
.5	.17
.6	.18
.7	.19
.8	.22
.9	.24
.95	.27
.975	.30
.99	.34
.999	.40
.9999	.45
.99999	.50
.999999	.55
mean	.173
var	.003

Thus, using the random-number-generated experience for n=18 from ref [1] and Mr. Venter directly, operational bootstrapping yields an expected least squares credibility complement $E[\frac{S^{*}_{..}}{18 \cdot T^{*}_{..}}]$ of .173. An adjustment described after Section 5 will suggest that we multiply the .173 by $\frac{18}{17}$, producing .183; whereas $\frac{s^2}{s^2+n \cdot t^2} = .224$.

Section 5. A METHOD TO TEST THE STATISTICAL EFFICIENCY OF THE (OPERATIONAL) BOOTSTRAP FOR THE LEAST SQUARES CREDIBILITY.

Even having a good algorithm to do generalized numerical convolutions, we still need an algorithm to handle implicitly and without random numbers the required sampling and resampling. I will now sketch how such an algorithm was (and can be) constructed.

Let f_{X_g} denote what is, in effect, the discretized representation of the distribution to which Risk g (g=1,2,...,N) is subject. Let f_{X_g} be represented in element notation as

$$\left[x1_j \; p1_j\right]_{j=1,2,\cdots,j_g}$$

where j_g is the number of lines in f_{X_g}.

Let n be the number of (possibly duplicate) items ultimately to be included in each of the samples and the resamples.

The calculational process will possess a Markov property, if we structure it as follows.

For any fixed Risk g process each of the xi_j in order by j $(j=1,2,\ldots,j_g)$, referring to value of j as a "stage" in the process. At stage j we shall use

$$_jf_{X^{**},Y^{**},V,V'}$$

to represent the 4-dimensional probability distribution of (X^{**},Y^{**},V,V') where

X^{**} is the resample mean where the resample is of size V',

Y^{**} is the resample variance where the resample of of size V',

V is the number of items included in the sample, and

V' is the number of items included in the resample,

If v_1 and v_2 (with $v_1<=v_2<=n$) are two values of V and if v'_1 and v'_2 (with $v'_1<=v'_2<=n$) are two values of V', then we can speak of the transition from the joint state (v_1,v'_2) at stage j to the joint state (v_2,v'_2) at stage j+1 where

v_2-v_1 is the number of duplicate items being added to the sample between stage j and stage j+1, and

$v'_2-v'_1$ is the number of duplicate items being added to the resample between stage j and stage j+1.

The calculational process involves using the 4-dimensional distribution

$_{j}f_{X^{**},Y^{**},V,V'}$, which we might relabel as

$$_{j}(\text{resample mean, resample variance})_{v_1,v_1'}$$

and

the amount $x1_{j+1}$ and probability $p1_{j+1}$ on the $j+1^{st}$ line of f_{X_g}

in a generalized numerical convolution to obtain the bivariate distribution

$$_{j+1}f_{X^{**},Y^{**},V_2,V_2'}$$

Actually, since the state (v_2,v_2') at stage j+1 can be arrived at from various states (v_1,v_1') at stage j, the complete distribution

$$_{j+1}f_{X^{**},Y^{**},V_2,V_2'}$$

is obtained by generating and merging together each of the resulting relevant partial bivariate distributions.

We eventually will have generated

$$_{j_1}f_{X^{**},Y^{**},V_2,n}$$

where V_2 runs from 1 to n;

and we merge such distributions together to obtain the 3-dimensional marginal distribution

$$j_g f_{X^{**},Y^{**},\cdot,n}.$$

Computer Note: Although X and Y are real random variables, V_2 *and* V_2' *take on only the nonnegative positive integers from 0 to n. Rather than record the distributions* $j_g f_{X^{**},Y^{**},V_2,V_2'}$ *as 4-dimensional distributions we* can (and do) *record for each joint state* (v_2,v_2') *the part of the 2-dimensional marginal distribution* $j_g f_{X^{**},Y^{**}}$ *which is associated with that joint state. The number of such joint states does not exceed* $n\cdot(n+1)/2$, *so this procedure is practical for n less than (say) 50. A separate disk file is created for the distribution associated with each joint state.*

Here are some hints on how the algorithm was (and can be) constructed.

Hint #1

Consider the first line of the distribution f_{X_g}, namely, $x1_1$ $p1_1$. We wish to enumerate the possible contributions of this line to the resample means and resample variances, where both samples and resamples will eventually be of size n.

The contribution of this line to the mean of a resample that will be adding exactly v_2' duplicates of such items will be simply $v_2' \cdot x1_1$. And the probability that both the sample will be adding exactly v_2 duplicates of such items and the resample will include exactly v_2' of such items can be calculated as

$$b_{v_2}(n;p1_1) \cdot b_{v_2'}(n;v_2/n)$$

where b indicates a binomial distribution defined by the two parameters inside the parentheses.

Hint #2

Suppose we are about to process line j of f_{X_g}, namely, $(x1_j\ p1_j)$. We wish to enumerate the possible contributions of this line to the resample means and resample variances. The probability that both the

sample will be adding exactly v_2-v_1 of such duplicate items and the resample will include exactly $v'_2-v'_1$ of such duplicate items can be calculated as

$$b_{v_2-v_1}(n-v_1; p1_j/(1-p1_1-p1_2-\cdots-p1_{j-1}))$$

$$\cdot$$

$$b_{v'_2-v'_1}(n-v'_1;\ (v_2-v_1)/(n-v_1))$$

where

v_2-v_1 is is number of duplicates of such items being added to the sample,

$n-v_1$ is the number of duplicates of such items which could be added to the sample,

$p1_j/(1-p1_1-p1_2-\cdots-p1_{j-1})$ is a conditional probability for the j^{th} line of f_{X_g} given that the previous j-1 lines of f_{X_g} have already been completely processed,

$v'_2-v'_1$ is the number of duplicates of such items being added to the resample,

$n-v'_1$ is the number of duplicates of such iesm which could added to the resample,

and **b** is as defined above.

Hint #3

When we process the last line of f_{X_g}, the number of $x1_{j_g}$'s about to be included in the resample must be equal to n-v' since the resample must finally contain exactly n items. Thus, in this situation the probability associated with that event is 1.

Hint #4

If at any stage in the calculational process the sample already contains n items, then the item currently being considered cannot contribute to the resample and the related probability would be zero.

Hint #5

At any stage in the calculational process an item cannot contribute to the resample if none of such items is being included in the sample.

Hint #6

As you begin to process a new line of $f_{X_{q}}$, some computer time can be saved by letting the computer determine some of the bivariate distributions which would produce zero or negligible total probability at the next stage.

Continuing The Numerical Example

Even for n as low as 18 the volume of calculations required, using the algorithm described in Section 5 to determine the bivariate distribution of ($X_{g\cdot}$, $S_{g\cdot}$) for even one value of g is unusually high. Calculating the distributions which yielded the statistics in this Section were performed on a Hewlitt Packard 720 computer at Washington National using a C-language equivalent of the above-sketched algorithm in the command-and-convolute package COCONUT™ available from MathWare. The run for one Risk took a full week to complete, even on such a fast computer. $E[X_{1\cdot}^{**}]$ and $Var[S_{1\cdot}^{**}]$ turned out to be .6452 and .3749, respectively. The .6452 agrees exactly with $E[X_1]$ and the .3749 compares with $Var[S_1] = .3974$. So, $\frac{Var[S_1^{**}]}{Var[S_1]} = .944$, indicating that $Var[S_{1\cdot}^{**}]$ is understating $Var[S_{1\cdot}]$

Assume for the moment that $\frac{Var[S_{i\cdot}^{**}]}{Var[S_{i\cdot}]}$ for each i=2,3,...,9 was also .944. Then would could also conclude that $\frac{E[S_{\cdot\cdot}^{**}]}{n \cdot E[T_{\cdot\cdot}^{**}]}$ will be .944 times $\frac{s^2}{s^2+n \cdot t^2}$.

To get an idea of how stable the ratio $\frac{Var[S_{\cdot\cdot}^{**}]}{Var[S_{\cdot\cdot}]}$ is likely to be as we move from distribution to distribution I calculated such ratios on my Gateway 486/33 (for various positive integer values of n) for the unit normal distribution on the one hand and the χ_{19} distribution on the other hand, with

the following results: the ratio $\frac{Var[S^{**}_{..}]}{Var[S_{..}]}$ for each of these distributions turned out to be almost exactly $\frac{n-1}{n}$, where n is the number of items in each of the samples and resamples. Note that $\frac{17}{18}$ is approximately equal to .944, which is the value we obtained above for a Gamma distribution with mean .6452 and variance .3974 and n=18.

In the next section Mr. Charles S. Fuhrer, FSA, of Washington National Insurance Company shows that the ratio $\frac{Var[S^{**}_{..}]}{Var[S_{..}]}$ is indeed always $\frac{n-1}{n}$ where the bootstrap distribution is being used to estimate the variance.

Thus, $E[\frac{S^{*}_{..}}{18 \cdot T^{*}_{..}}]$ = .173 (calculated by bootstrapping in Section 4) should be divided by $\frac{17}{18}$, producing .183, to try to obtain an unbiased estimate of $\frac{s^2}{s^2+n \cdot t^2}$. Of course, the .173 (and hence the .183) is based on Mr. Venter's particular 3 Random-Number-Generated Trials of 6 Years Combined Into One Trial fo 18 Years,

shown in a table in Section 4. Further runs could be performed to calculate the distribution of $(X^{**}_{g.}, S^{**}_{g.})$ for g=2,3,...,9; and, we could then generate the trivariate distributions described in Section 3, obtain a distribution of $\frac{S^{**}_{..}}{n \cdot T^{**}_{..}}$ nd draw further conclusions about any further bias inherent in using $\frac{S_{..}}{n \cdot T_{..}}$ as an estimator for the complement of the least squares credibility.

Appendix. DERIVATION OF THE FACTOR $\frac{n-1}{n}$

(The symbols in this Appendix differ from the symbols used in the body of the paper.)

Fix $x_1, x_2, \ldots, x_n$.

Let $Y_1, Y_2, \ldots, Y_m$ be distributed discretely with

$$\Pr\{Y_j = x_i\} = \frac{1}{n} \quad (i=1,2,\ldots,n).$$

Let

$$S_Y^2 = \frac{\sum_{j=1}^{m} (Y_j - \bar{Y})^2}{m} \quad \text{and} \quad \bar{Y} = \frac{\sum_{j=1}^{m} Y_j}{m}.$$

Let $S_X^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$ and $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

It is easy to show that

$$E[S_Y^2] = (\frac{m-1}{m})S_X^2.$$

Now let $\{X_1, X_2, \ldots, X_n\}$ be iid random variables distributed as X

with

$$E[X_i]=\mu \text{ and } V[X_i]=\sigma^2.$$

Then the mean of the resample variances is

$$E[S_Y^2] = E_X[E_Y[S_Y^2|X]] = E_X[(\frac{m-1}{m})S_X^2] = (\frac{n-1}{n})(\frac{m-1}{m})\sigma^2.$$

Now consider the variance $V[\bar{Y}]$ of the resample means $\bar{Y}$.

$$V[\bar{Y}] = V_X[E_Y[\bar{Y}|X]] + E_X[V_Y[\bar{Y}|X]]$$

$$= V_X[\bar{X}] + E_X[\frac{S_x^2}{m}]$$

$$= \frac{\sigma^2}{n} + \frac{(n-1)}{nm}\cdot\sigma^2$$

$$= (\frac{1}{n} + \frac{(n-1)}{nm})\cdot\sigma^2$$

$$= \frac{\sigma^2}{n}\cdot(1+\frac{n-1}{m})$$

or, if $m=n$,

$$= \sigma^2\cdot\frac{(2n-1)}{n^2}.$$

REFERENCES

Bailey, W. A. (1993) Six Bridges to Ψ's. *ARCH, Vol 1993.1.*

Venter, G. G. (1990) Credibility. *In Foundations of Casualty Actuarial Science*, pp. 416-445. Casualty Actuarial Society, New York.