Simulation Based Estimation: A Case Study in Oncology (SIMEST) and A Case Study in Portfolio Selection (SIMUGRAM)

James R. Thompson
Dept. Statistics, Rice University
Houston, Texas 77251-1892
email:thomp@rice.edu

ABSTRACT. Since the time of Poisson, stochastic processes have been axiomitazed in the temporally forward direction. Yet for nearly a century, estimation of parameters and even forecasts have been based on likelihood approaches which start with temporally indexed data and then look backwards in time. I shall be using the philosophy of Karl Pearson [1] in this paper where I create, using a forwards model, a large virtual universe of happenings based on the assumption of four parameters characterizing an oncological process bassed on four Poissonian processes. Bins will be formed in the real time space based on the actual data of s real world system of times of discovery of primary and secondary tumors and use bin boundaries which enclose roughly 5% of the actual tumor discover data and compare the bin proportions of virtual data with the proportions of actual data in each of the bins. This will enable us to use Karl Pearson's goodness of fit criterion as the objective function for a Nelder-Mead optimization. We present here an oncological example where the objective is to estimate four parameters relevant to the progression of beast cancer. This procedure I named the SIMEST paradigm.

Then we briefly describe the patented SIMUGRAM for estimating the distribution of portfolio values using daily resampling strategies. This procedure makes minimal model assumptions and is completely data based.

Keywords: breast cancer, forward estimation, portfolio forecasting

1 SIMEST for Parameter Estimation Using Breast Cancer Data

Let us consider the practical estimation of growth parameters using a data set of 116 women who presented with primary breast cancer at the Curie-Sklodowska Cancer Center of Warsaw . The first approach attempted by Thompson, Brown and Bartoszyński [2] was a maximum likelihood approach. The time required to grind through the mathematics of the likelihood function and its programming took around 1.5 person years by Thompson working on a Polish sabbatical with Bar-

toszyński. We obtained useful results, such as the fact that many of the secondary tumors were not generated by metastasis, but rather by systemic generation. One thing one always expects from a model-based approach is that, once the relevant parameters have been estimated, many things one had not planned to look for can be found. For example, tumor doubling time is 2.2 months. The median time from primary origination to detection is 59.2 months and at this time the tumor consists of 9.3×10^7 cells. However the likelihood approach was simply not computationally practical. The problem is that the axioms go in the forward temporal direction whereas the likelhood looks backwards.

Upon returning to Houston, it occurred to Thompson [3−5] that one might use simulation as a means of creating thousands of pseudo tumors based on parameter assumptions and comparing their binned detection times with the clinical data using a simple goodness of fit measure. Then one might seek estimates for the parameters by minimizing this criterion function.

The SIMEST Paradigm

First, we observe how the forward approach enables us to eliminate those hypotheses which were, essentially a practical necessity if a likelihood function was to be obtained. Our new axioms are simply:

Hypothesis 1. For any patient, each tumor originates from a single cell and grows at exponential rate α.

Hypothesis 2. The probability that the primary tumor will be detected and removed in $[t, t + \Delta t)$ is given by $bY_0(t)\Delta t + o(\Delta t)$. The probability that a tumor of size $Y(t)$ will be detected in $[t, t + \Delta t)$ is given by $bY(t)\Delta t + o(\Delta t)$.

Hypothesis 3. The probability of a metastasis in $[t, t + \Delta t)$ is $a\Delta t\times$ (total tumor mass present).

Hypothesis 4. The probability of a systemic occurrence of a tumor in $[t, t + \Delta t)$ equals $\lambda\Delta t + o(\Delta t)$, independent of the prior history of the patient.

In order to simulate, for a given value of (α, a, b, λ), a quasidata set of secondary tumors, we must first define:

t_D = time of detection of primary tumor;
t_M = time of origin of first metastasis;
t_S = time of origin of first systemic tumor;
t_R = time of origin of first recurrent tumor;
t_d = time from t_R to detection of first recurrent tumor;
t_{dM} = time to detection of first metastasis.
t_{DR} = time from t_D to detection of first recurrent tumor.

We proceed with generating all the times mentioned above, using the fact that the distribution of a continuous random variable is uniform over the unit interval. For example, generating a random number u from the uniform distribution on the unit interval, if $F(\cdot)$ is the appropriate cumulative distribution function for a time, t, we set $t = F^{-1}(u)$. Then, assuming that the tumor volume at time t is

$$v(t) = ce^{\alpha t}, \text{ where } c \text{ is the volume of one cell,} \tag{1}$$

we have

$$F_M(t) = 1 - \exp\left(-\frac{ac}{\alpha}e^{\alpha t_M}\right). \tag{2}$$

Similarly, we have

$$\begin{aligned} F_D(t_D) &= 1 - \exp\left(-\int_0^{t_D} bce^{\alpha\tau}d\tau\right) \\ &= 1 - \exp\left(-\frac{bc}{\alpha}e^{\alpha t_D}\right), \end{aligned} \tag{3}$$

$$F_S = 1 - e^{-\lambda t_S}, \tag{4}$$

and

$$F_d(t_d) = 1 - \exp\left(-\frac{bc}{\alpha}e^{\alpha t_d}\right). \tag{5}$$

Equations (2)−(5) represent the four Poissonian processes charactrizing my model. Using the actual times of discovery of secondary tumors $t_1 \leq t_2 \leq, \ldots, \leq t_n$, we generate k bins (k is a bandwidth parameter, typically chosen so that each bin contains 5% of the t values. In actual tumor situations, because of recording protocols, we may not be able to put the same number of secondary tumors in each bin. Let us suppose that the observed proportions are given by $(p_1, p_2, \ldots, p_k)$. We shall generate N recurrences $s_1 < s_2 < \ldots < s_N$. The observed proportions in each of the bins will be denoted $\pi_1, \pi_2, \ldots, \pi_k$. The goodness of fit corresponding to (α, λ, a, b) will be given by

$$\chi^2(\alpha, \lambda, a, b) = \sum_{j=1}^{k} \frac{(\pi_j(\alpha, \lambda, a, b) - p_j)^2}{\pi_j(\alpha, \lambda, a, b)}. \tag{6}$$

As a practical matter, we may replace $\pi_j(\alpha, \lambda, a, b)$ by p_j, since with (α, λ, a, b) far away from truth, $\pi_j(\alpha, \lambda, a, b)$ may well be zero. Then the following algorithm generates the times of detection of quasisecondary tumors for the particular parameter value (α, λ, a, b).

Secondary Tumor Simulation (α, λ, a, b)

Generate t_D
$j = 0$
$i = 0$
Repeat until $t_M(j) > t_D$
$j = j + 1$
Generate $t_M(j)$
Generate $t_{dM}(j)$
$t_{dM}(j) \leftarrow t_{dM}(j) + t_M(j)$
If $t_{dM}(j) < t_D$, then $t_{dM}(j) \leftarrow \infty$
Repeat until $t_S > 10t_D$
$i = i + 1$
Generate $t_{dS}(i)$
$t_{dS}(i) \leftarrow t_{dS}(i) + t_S(i)$
$s \leftarrow \min [t_{dM}(j), t_{dS}(i)]$
Return s
End Repeat

This algorithm does still have some simplifying assumptions. For example, we assume that metastases of metastases will probably not be detected before the metastases themselves. We assume that the primary will be detected before a metastasis, and so on. Note, however, that the algorithm uses much less restrictive simplifying assumptions than those that led to the terms of the closed-form likelihood. Even more importantly, the Secondary Tumor Simulation algorithm can be discerned in a few minutes, whereas a likelihood argument is frequently the work of months.

Another advantage of the forward simulation approach is its ease of modification. Those who are familiar with "backward" approaches based on the likelihood or the moment generating function are only too familiar with the experience of a slight modification causing the investigator to go back to the start and begin anew. This is again a consequence of the tangles required to be examined if a backward approach is used. However, a modification of the axioms generally causes slight inconvenience to the forward simulator.

For example, we might add:

Hypothesis 5. A fraction γ of the patients ceases to be at systemic risk at the time of removal of the primary tumor if no secondary tumors exist at that time. A fraction $1 - \gamma$ of the patients remains at systemic risk throughout their lives.

Clearly, adding Hypothesis 5 will cause considerable work if we insist on using the classical aggregation approach of maximum likelihood. However, in the for-

ward simulation method we simply add the following lines to the Secondary Tumor Simulation code:

> Generate u from $U(0,1)$
> If $u > \gamma$, then proceed as in the Secondary Tumor Simulation code
> If $u < \gamma$, then proceed as in the Secondary Tumor Simulation code except replace the step "Repeat until $t_S > 10t_D$" with the step "Repeat until $t_S(i) > t_D$."

Thompson presented his flowchart to E. Neely Atkinson, then a graduate student, and with 5 person hours of work he arrived at the same parameter estimates as those obtained by the tedious likelihood approach [2]. Thompson now gives the Polish breast cancer problem to students in his simulation course.

In the discussion of metastasis and systemic occurrence of secondary tumors, we have used a model supported by data to try to gain some insight into a part of the complexities of the progression of cancer in a patient. Perhaps this sort of approach should be termed *speculative data analysis.* In the current example, we were guided by a nonparametric intensity function estimate [2], which was surprisingly nonincreasing, to conjecture a model, which enabled us to test systemic origin against metastatic origin on something like a level playing field. The fit without the systemic term was so bad that anything like a comparison of goodness-of-fit statistics was unnecessary.

It is interesting to note that the implementation of SIMEST is generally faster on the computer than working through the estimation with the closed-form likelihood. In the four-parameter oncological example we have considered here, the running time of SIMEST was 10% of the likelihood approach.

2 A Glimpse at The Patented SIMUGRAM Paradigm for Portfolio Optimization

The use of stochastic modeling in the management of portfolios is of some antiquity, going back at least to Bernoulli's work in 1738 [6]. Over the succeeding years, many strategies for investing in the stock market have been considered. Some of these are delineated by Bernstein [7] and Thompson et alia [8]. Findlay et alia [9] note that no model is good if it does not confirm to data from the real world. For some years, the neoclassical school frequently associated with the University of Chicago

has been particularly popular, some of it is based on the *Efficient Market Hypothesis* which is roughly based on the martingale assumption that for any stock $S(t)$, the expected value at some future time $t + u$ will be equal to $S(t)$. All characteristics of a stock are claimed to be included in its current price. This theory has many consequences. The Capital Market Line of Sharpe is a line, drawn on a growth versus standard deviation curve starting on the left with the zero volatility of a T−Bill with zero standard deviation to the growth rate and standard deviation of a market cap weighted portfolio using all publicly traded stocks. According to devotees of the Efficient Market Hypothesis there are no stocks or portfolios of stocks which can lie above this CML Any other strategy must give a portfolio lying below and to the right of the CML. On the basis of this theory, we have a plethora of market cap weighted index funds, ranging from total market funds, to spider funds based on a combination of securities within a specific sector. For his development of Capital Market Theory in 1964 [10], William Sharpe received, in 1990, a Nobel Prize in Economics. This theory is an underpining of the Index Fund Theory of John Bogle [11], which holds that one should invest, for example, in a market cap weigted portfolio from the S&P 500. This has become the basis of many mutual funds used for the purpose of Individual Retirement Accounts (IRA).

However, Thompson et alia (2006) [12] have considered looking, year by year, at 50,000 portfolios consisting of random selections of stocks from the 1,000 highest market cap securities. For the 33 years from 1970 through 2002, not simply a flukeish few, but a staggering 65% of the portfolios selected randomly from the 1,000 largest market cap stocks lie above the CML. So, now we have empirical evidence to the effect that index funds really do not have some sort of cosmic connection to optimality. If we can beat the index with a randomly selected portfolio, that does not indicate that we should use random selection as our new evangel for portfolio design. Rather, we have simply put to rest the notion of the intrinsic optimality of index funds. We now know that there is hope for market analysis which swims against the tide of those who claim there is no possible portfolio selection which gives returns above the CML. The natural option to consider when random selection beats a standard procedure is to use equal weights on all stocks in the universe considered. We shall demonstrate this result shortly. Further, we might actually try and find a robust rule for optimal selection of weights in a portfolio of stocks. This was the objective in Thompson's patented algorithm sometimes referred to as the SIMUGRAM [13].

Let us suppose we have a data base showing the year to year change in a stock $S(t)$ or a stock index. We can then obtain a data base of terms like

$$R_i = \log(\frac{S(t_i)}{S(t_{i-1})}). \tag{7}$$

In other words, we know that

$$S(t_i) = S(t_{i-1}) \times \exp(R_i). \tag{8}$$

Suppose we have a data base of n such terms, $\{R_1, R_2, \ldots, R_n\}$. Let us make the (frequently reasonable) assumption that the ups and downs of the stock or the index in the past are a good guide to the ups and downs in the future. It would not be a good idea, if we wished to forecast the value of the stock five years in advance, randomly to sample (with replacement) five of the R_i's, say, $\{R_3, R_{17}, R_{20}, R_{20}, R_{31}\}$ and use

$$\hat{S}(5) = S(0) \times \exp[R_3 + R_{17} + R_{20} + R_{20} + R_{31}].$$

On the other hand, if we wished to obtain, not a point estimate for $S(T)$, but an estimate for the distribution of possible values of $S(5)$, experience shows that this frequently can be done as follows:

Portfolio Resampling

1. Enter $S(0)$, T, and the $\{R_i\}$.
2. Repeat 10,000 times
3. For pass i, randomly sample with replacement T values from $\{R_1, R_2, \ldots, R_n\}$, say, $\{R_{i1}, R_{i2}, R_{i3}, R_{i4}, \ldots\}$.
4. Compute
$$SS(T) = S(0) \times \exp[R_{i1} + R_{i2} + R_{i3} + R_{i4} + \ldots]$$
Clearly we use T resampled R values to obtain.
5. Obtain the empirical cumulative distribution function from the resulting 10,000 values of $SS(T)$. That is, compute
$$F_T(v) = \frac{\text{Number of sorted values } \{SS(T)\} \leq v}{10,000}. \tag{10}$$

By looking at the simulations for, say, five, ten, twenty and forty years in the future, an investor can examine the historically based outcomes of buying a security in the light of his/her anticipated needs. Of ccourse, the farther into the future we extrapolate, the more dubious will be the stationarity assumptions of first order stationarity across the ensemble of stocks.

2.1 The Multivariate Case

Here, for each of p stocks, we compute

$$R_{i,j} = \log(\frac{S(t_{i,j})}{S(t_{i,j-1})}). \tag{11}$$

Multivariate Portfolio Resampling

We then proceed very much as in the parametric case:

1. Enter $\{S_j(0)\}_{j=1}^k$, T, and the $\{R_{i,j}\}_{i,j}$.
2. Repeat 10,000 times
3. For pass i, randomly sample with replacement T values from the length of the historical list, say, $l_{i,1}, l_{i,2}, \ldots, l_{i,T}$.
4. For each stock j

$$SS_{i,j}(T) = S_j(0) \times \exp[R_{l_{(i,1)},j} + R_{l_{(i,2)},j} + \ldots + R_{l_{(i,T)},j}]$$

5. Store $(SS_{i,1}, SS_{i,2}, \ldots, SS_{i,j})$ as a row vector $\mathbf{SS}_i$.
6. End Repeat

Now for a portfolio of p stocks, balanced according to

$$P(t) = \sum_{i=1}^{p} c_i S_i(t), \tag{12}$$

where the weights are non-negative and sum to one, we simply use the,

An Algorithm for Resampling Portfolios

1. Enter $\mathbf{SS}(T)$, $\{c_j\}$.
2. For all i from 1 to 10,000, find $P_i(T) = \sum_{j=1}^{p} c_j SS_{i,j}(T)$.
3. Sort the $P_i(T)$.
4. $F(v) = \frac{\text{Number of } P_i(T) \leq v}{10{,}000}$.

We may then proceed to obtain simulations of the portfolio value at a given time. Because of the correlations between stock values, it is essential that, when we randomly select a year, we sample the annual growth factors of the stocks in the portfolio for that year. In Figure 1, we show what would have happened had we used an equal weighting strategy using stocks from the S&P 100, rebalancing the portfolio once a year, during the years 1970-2002. The aggregate return is equivalent to continuously compounded interest rate of 13.2%. This compares to an S&P 100 return of 8.4%. (Both returns are exclusive of dividends.) It is also to be noted that over the 35 years, the total negative returns in losing years are a minus 90.57% for the equal weight portfolio, as opposed to a minus 118.13% for the S&P 100 index fund. We have no immediate explanation for the superior performance of the equal weight portfolio. There are, as we have seen, excellent theoretical explanations for why the S&P 100 should be superior to non market cap weighted funds. Unfortunately, here theory does not conform to facts. Our general recommendation when theory does not conform to facts is to try to develop, ultimately, an alternative theory. For the short to medium term, we should develop strategies based on rules developed empirically from data.

One thing we might try is to build a "equal weight fund" with equal amounts of capital invested in each security in the portfolio. (Perhaps the weighting by market cap penalizes the portfolio for investing too much in large companies.) In Figure 1, we show what would have happened had we used an equal weighting strategy using stocks from the S&P 100, rebalancing the portfolio once a year, during the years 1970-2002. It is also to be noted that over the 35 years, the total negative returns in losing years are a minus 90.57% for the equal weight portfolio, as opposed to a minus 118.13% for the S&P 100 index fund. We have no immediate explanation for the superior performance of the equal weight portfolio. There are excellent theoretical explanations for why the market cap weighted S&P 100 should be superior to otherly weighted funds. Unfortunately, here theory does not conform to facts. Our general recommendation when theory does not conform to facts is to try to develop, ultimately, an alternative theory. For the short to medium term, we should develop strategies based on rules developed empirically from data.

Other strategies might be based on rules much more complicated than that of equal weighting. One such is the patented Simugram. portfolio paradigm of Thompson [5]. This is a computer intensive expert system which looks at the synchronized historical performance of the stocks in a selection set and uses this information to develop a high return, low risk portfolio. Again, the portfolio is rebalanced once a year. (The Simugram algorithm does not generally conform to a "buy and hold" strategy.) We show the results of 35 years of applying the Simugram portfolio paradigm, ex ante, to the stocks in the S&P 100 in Figure 1. We note that the ag-

gregate return, exclusive of dividends, is equivalent to a continuously compounded interest rate of 20.0%. This is a rate of return comparable to those generally associated with Warren Buffett's Berkshire Hathaway. We note that, over the 35 years starting in 1970, the total negative returns in losing years are a minus 112.74%, less than those experienced by the market cap weighted S&P 100 but more than those of the equally weighted portfolio.

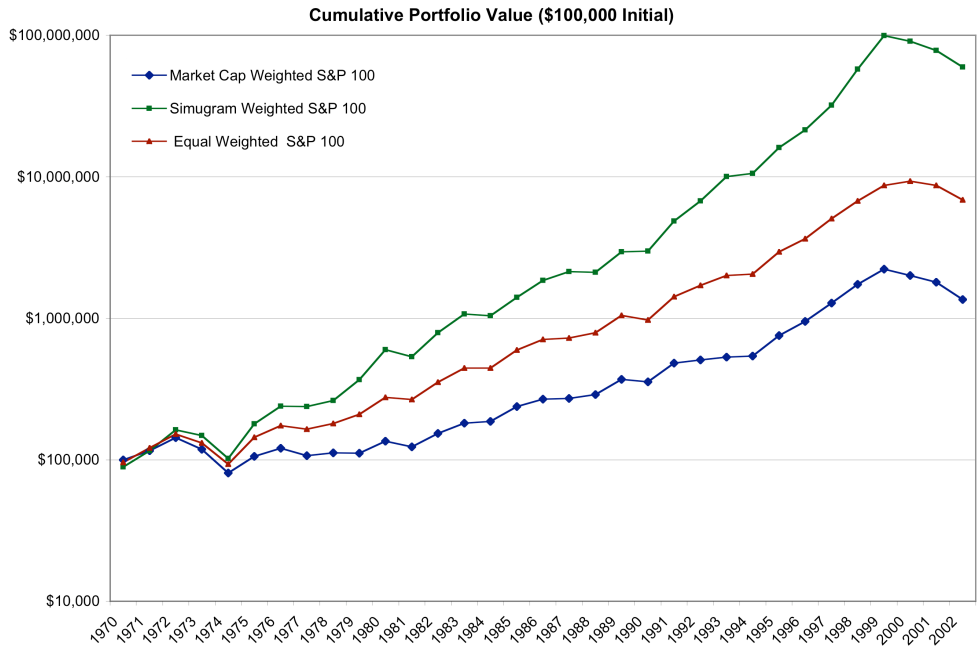

Figure 1. Comparison of SIMUGRAM S&P 100 with S&P 100 Index.

Obviously neither the SIMUGRAM nor any other portfolio buying rule is free of risk. The Iraq War seriously damaged what efficiency there was in the market and after 2009, both SIMUGRAM and the rules of Berkshire−Hathaway failed, for a long time, to outperform the simple market cap weighted index.

The Equal Weight option is closely emulated by the MaxMedian Rule which we show below

The MaxMedian Rule

1. **For the 500 stocks in the S&P 500 look back at the daily returns S(j,t) for the preceding year the day to day ratios r(j,t) = S(j,t)/S(j,t−1).**
2. **Sort these for the year's trading days.**
3. **Discard all r values equal to one. in the 500 medians of the ratios.**
4. **Invest equally in the 20 stocks with the largest medians.**
5. **Hold for one year and then liquidate.**

MaxMedian has the advantage of enabling one to form his or her own portfolio with a market utility very near that of the Equal Weight Rule. It requires the purchase of *hquote* software for a one time fee of $80. All computations may be done on an Excel spreadsheet.

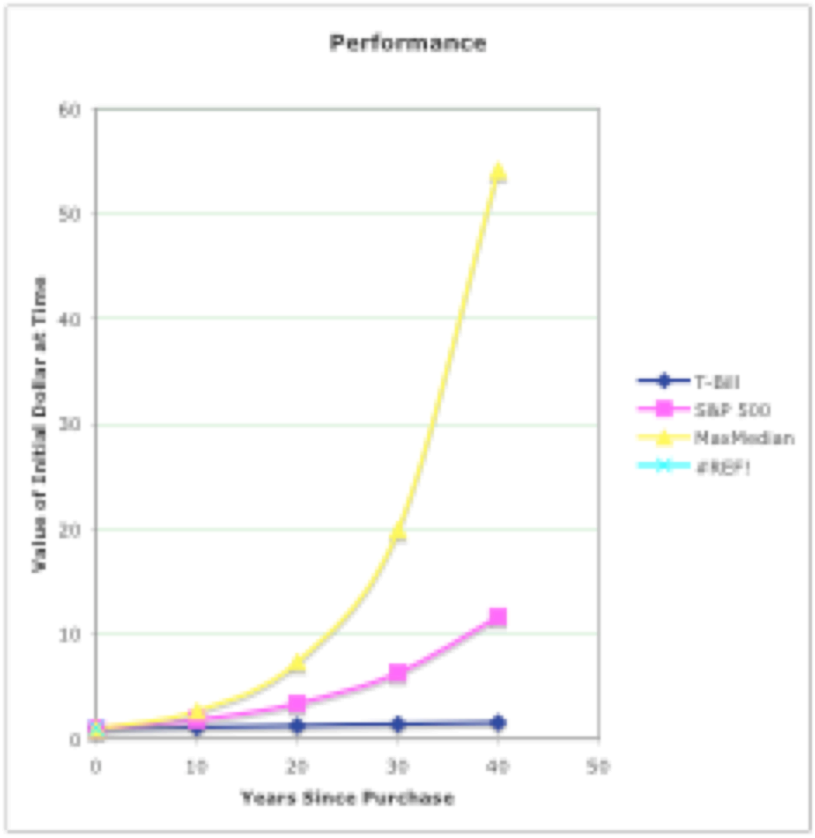

Figure 2. Comparison of Equal Weight, MaxMedian Weight, and Market Cap Weight

References

[1] Scott, David W., Tapia Richard A. and Thompson, James R. (1978)."Karl Pearson Was Right," *Computer Science and Statistics: Tenth Annual Symposium on the Interface*, 179-183.

[2] Bartoszyński, Robert, Brown, Barry W., and Thompson, James R. (1982). "Metastatic and systemic factors in neoplastic progression," in *Probability Models and Cancer*, L. LeCam and J. Neyman, eds. New York: North-Holland., 253−264.

[3] Bartoszyński,Robert, Brown, Barry W., McBride, Charles, and Thompson, James R. (1981). "Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary Poisson process," Annals of Statistics, v. 9, pp. 1050−1060.

[4] Thompson, James R., Atkinson, E. Neely, and Brown, Barry W. (1987). "SIMEST: an algorithm for simulation based estimation of parameters characterizing a stochastic process," in *Cancer Modeling*. Thompson, J. and Brown, B., eds. New York: Marcel Dekker, 387−415.

[5] Thompson, James R. (1989). *Empirical Model Building*. New York: John Wiley & Sons, 125−131.

[6] Bernoulli, Daniel (1954), Exposition of a new theory on the measurement of risk, (Louise Sommer, trans.). Econometrica, January, 23-36 (first published in 1738).

[7] Bernstein, Peter L. (1996). *Against the Gods: The Remarkable Story of Risk*. New York: John Wiley & Sons, 99−115.

[8] Thompson, James R., Williams, Edward E. and Findlay, M Chapman III (2003). *Models for Investors in Real World Markets*. Hoboken, N.J.: John Wiley& Sons.

[9] Findlay, M. Chapman, Williams, Edward E., and Thompson, James R. (2003)."Why we all held our breath when the market reopened," *Journal of Portfolio Management*, Spring, 91−100.

[10] Sharpe, William F. (1964)."Capital Asset Prices: A theory of market equilibrium under conditions of risk,"*Journal of Finance*, September, 425−442.

[11] Bogle, J. C. (1999). Common Sense and Mutual Funds: New Imperatives for the Intelligent Investor. New York: John Wiley & Sons.

[12] Thompson, James R., Baggett,L.Scott, Wojciechowski,William C. and Williams, Edward E. (2006). " Nobels for Nonsense", *The Journal of Post Keynesian Economics*, Fall, 3−18.

[13] Thompson, James R., US Patent 7,720,738 B2 May 18, (2010), *Methods and Apparatus for Determining A Return Distribution for An Investment Portfolio.*

[14] Thompson, James R., (20011), *Data Models and Reality*, Hoboken, N.J.: John Wiley& Sons 358−360.