Some Uses of the Freeware, R, for Studying Financial Phenomena

Larry Weldon
weldon@sfu.ca
Simon Fraser University

It is well-known that the computing revolution has changed the way statistics is practiced. Methods like regression analysis and time series analysis have been extended so much that the modern theory has little resemblance to the pre-computer theory: a primary concern used to be whether the design matrix could be inverted in a reasonable amount of time, or whether a graphical presentation of data could be prepared in time for a presentation. While modern statistics has learned to make use of computing's speed of computation, it is still limited by the pre-computer habit of focusing on the estimation and testing of parametric models. A major reason for the focus on parameters was the need to summarize a complex data set with a few numbers. But there are alternatives now: nonparametric smoothing along with graphical display is now an excellent method for extracting information from data, and simulation is an excellent way to study the reliability of signals extracted from data. Put another way, nonparametric smoothing is replacing parametric estimation, and simulation is replacing parametric hypothesis testing.

In this talk, I want to demonstrate these trends using some powerful, flexible, and free software called "R" (R Development Core Team (2008)). In the realm of financial phenomena, I have chosen the particular areas as follows:

Part I: A model of the equity-bond differential and long-term risk
Part II: The effects of random variation in mutual fund performance
Part III: The advantage of size for simple businesses
Part IV: A simulation model of health status & hospital utilization

Part IV is joint work with Leonard MacLean (Dalhousie) and Andrej Blejec (Ljubljana).

Before launching into these demonstrations, I want to comment on the utility of the simulation model approach to the study of financial matters. Many financial processes, such as stock markets, or the insurance businesses, are complex and experience ever-changing environments, so that the search for the ultimate model is a utopian quest. But if we can build a flexible simulation model that, with appropriate calibration, mimics current data, then we have a chance of judging the sensitivity to changing environments. Moreover, using repeated runs of the model, we can observe the variability due to unexplained factors, and learn to appreciate the degree to which the system is unpredictable. Simulation models can produce useful information, since they focus our attention on aspects of phenomena that can be predicted or controlled, and reveal those aspects of phenomena that cannot be predicted or controlled. "Modern computation and simulation" is identified by Stephen Stigler (2008) as one of the top ten ideas in the history of statistics.

As a pension fund trustee at SFU, I was frequently exposed to the "risk-reward" curve and the talk of "how much risk can you accept?". The curve was really a variability-reward curve and the variability was always short term (months or quarters). This characterization of

the investor's dilemma annoyed me since risk and variability are very different things. Risk is the chance of loss while variability is change over time. One could argue that defining risk as short-term variability is fair in a world of jargon, but I think it is misleading. Pension investors are almost always long term investors, and short term variability is of little consequence to them, or at least it should be so. If funds are accumulated over 40 years and spent over another 20 years, as in pension investments, then this is definitely a long term horizon. And yet investors are advised to reduce short term variability so they can "sleep at night". While long term investors sometimes do worry about short term variability, pension investment advisors should not buy into this tradition as if it were wise advice. Even lifetime funds instruments that change mix as investors' age are too focused on short-term variability.

Part I: A model of the equity-bond differential and long-term risk

A common method for reducing short-term variability is to hold bonds. For the long term investor, does it make sense to hold bonds to reduce short-term variability when it would usually reduce long term profitability? The crucial question is: How likely, and by how much, will long term investments in bonds underperform equities? One approach to this problem is to construct a model of the market for bonds and one for equities and run it for an extended period, twenty-five years say, many times. Of course, for the model to be useful, it must be calibrated to known data and also it must reproduce the characteristics of the respective markets.

The data based evidence we use is as follows: long term ROR for equities in Canada, using S&P/TSX as a proxy, is 10 percent per annum, and for bonds, using long bonds as a proxy, is about 7 percent per annum. And based on daily data from S&P/TSX, the mean of the absolute daily change for equities is about .60, and the SD of the absolute daily change is .67 percent. We also note that when larger "shocks" occur, they tend to be in the 3-5 percent range and only once in 25 years would they approach 10 percent. Our models incorporate this information.

We use a gamma distribution with mean .60 and SD .67 to simulate this absolute change in equities. The sign of the daily change in equities is calibrated to .53 for positive and .47 for negative daily changes, and day-to-day independence is assumed. We allow bonds to have the same stochastic model except that its moves are, on average, .675 as large as those of equities. Following Consigli et al (2009) we also introduce a shock system for equities that depends on the equity-bond daily ROR differential. In fact, the difference in the daily ROR for equities and bonds is the probability of a shock occurring, and the direction of the shock is determined by the sign of the difference. If equities are returning more than bonds, the shock, if it occurs, will be positive for equities, and if bonds are returning more than equities, the shock, if it occurs, will be negative for equities. With these parameters, the observed data (on average, 10 percent pa for equities and 7 percent pa for bonds, over 28 years) is reproduced. Moreover, the distribution of shocks, in percent of the equity index, both positive and negative, has been chosen as exponential with mean 1.67: most shocks are less than 2 percent but in 25 years, there would usually be 5-10 shocks greater than 2 percent, and the largest of these are usually in the range 5-10 percent. This seems to fit with the past 25 year experience of the S&P/TSX.

Fig 1 shows a typical simulation. Experience with the simulator produces outcomes similar to the observed series: Fig 2 shows the real data series for equities.

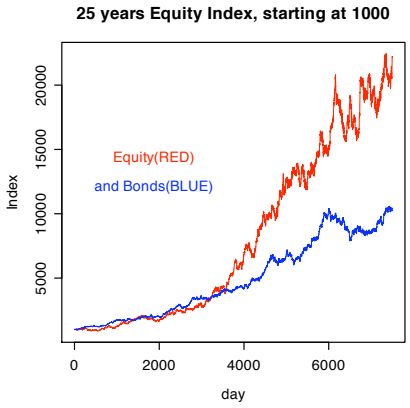
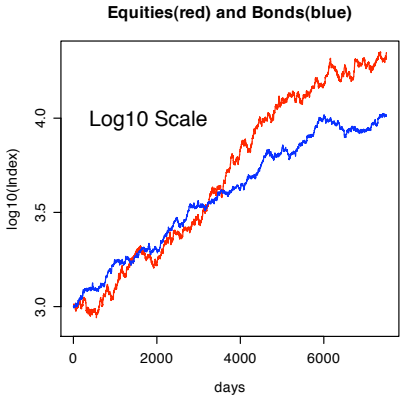
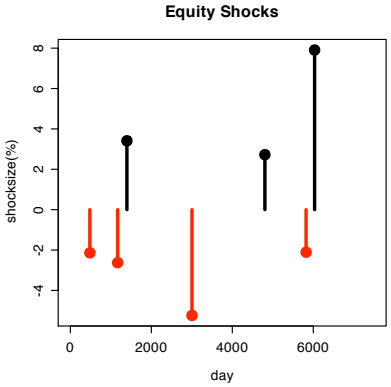


Fig 1:
These are typical outcomes of the simulation. The **first** graph shows the raw index, with equities ahead of bonds, and absolute variability increasing with level. The constant percentage variability is demonstrated in the **second** logarithmic graph. The **third** graph shows the equity shocks, positive and negative.





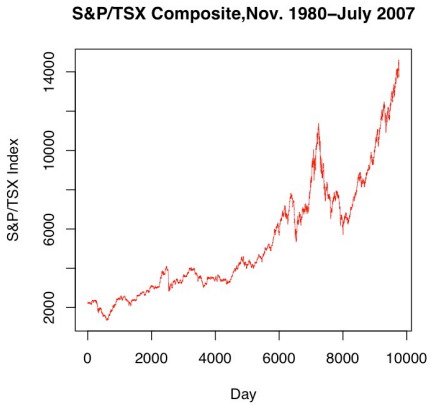
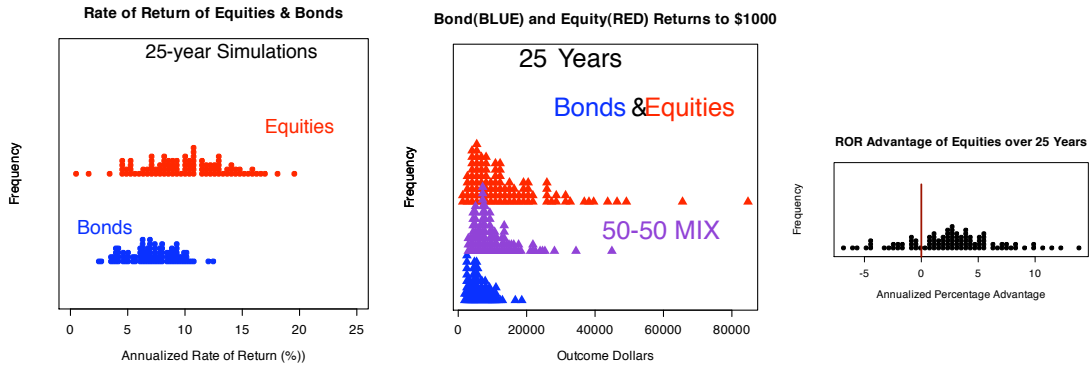


Fig 2: This is the real S&P TSX composite data for 1980-2007.

Once we are satisfied that the model is calibrated satisfactorily, what use can be made of the model? One item of interest is to compare the outcome for a long term investor of an investment in equities, or else in bonds, or in a 50-50 mix. Here is an example of 100 replications of the 25 year simulation of the model:

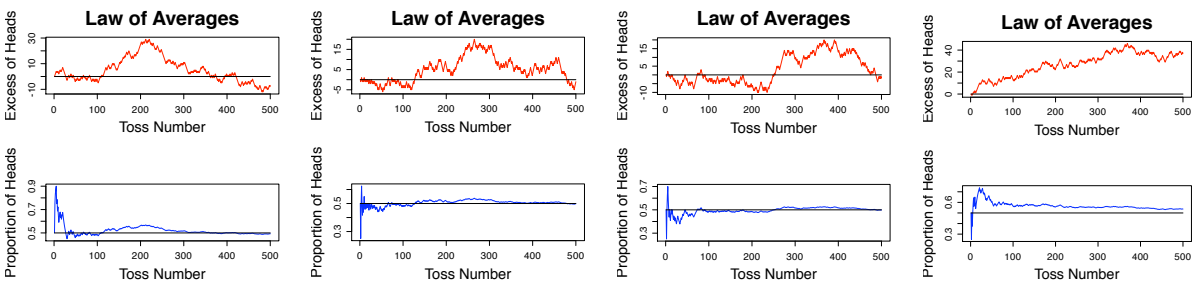
Fig 3: Bonds vs Equities



In this particular run the ROR for equities was 10.0% pa and for bonds was 7.15% pa and the pure equities portfolio produced higher returns in 77 of the 100 simulations. These results are typical. Although the bond and equity ROR distributions seem to have a large overlap, the pairwise comparison favors equities often. In a larger simulation, equity ROR exceeded bond ROR for this 25-year duration in 77.8 % of the simulations.

Now consider the investor who is trying to decide on the amount of "risk" he or she can accept in the quest for a good return. The decision could be more rationally made in view of the above graphs than selecting a point on the traditional "risk-reward" curve, since the latter is based on short term variability.

Note that even though equities outperform bonds on average, it can take a long time for this advantage to be realized. Even after 25 years, there is still a 22% chance that the bonds would have been better. This may be surprising – its appreciation requires an intuitive familiarity with random walks. Take for example the symmetric random walk, with steps of +1 or -1, each with probability 0.5. We know that the accumulation of displacements cannot have any trend useful for prediction, since the best prediction of future values is its expected value, and the expected value in the future is just the most recently observed value. However, look at typical outcomes of this random walk in Fig 4.



The red line is the symmetric random walk though may be surprising that it seems to embody various trends. The blue line just confirms that the law of large numbers is still true! This surprising occurrence of apparent trends is very useful to keep in mind in examining stock index trends. Morever, the long term variability of equity and bond returns is better appreciated with this random walk phenomenon in mind.

Part II: The effects of random variation in mutual fund performance

In this section, I will use a slightly simpler random walk for a specific purpose. The walk is supposed to mimic a fund of index on a daily basis. The steps are not symmetric: the probability p of a positive step is about .55. The size of the steps are fixed at 1/3 of 1 percent of the current value. It turns out that the rate of return over a long period with this model will mimic a poor fund if the p=.54, an average fund if p=.55, and a superior fund if p=.56. Here is an example of this effect:

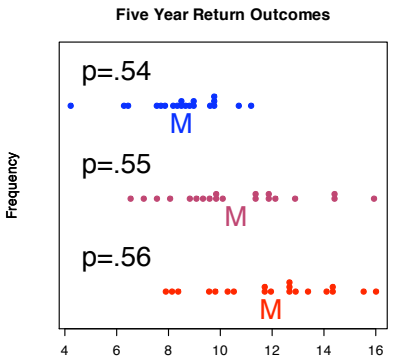


Fig 5: This graph shows how the fund manager's performance depends on the "quality" parameter p. p is the probability of a daily increase in fund value.

Now consider a group of 100 mutual fund investors who have personal p-values defined at random: the 100 p-values are designed to have mean .55 and symmetrical deviations from that mean are exponential with mean .005. So we expect a range of p values mostly in the range .53 to .57. Now we run these managers experiences for five years, and based on their rates of return we select the best 15% of managers. Then we look at the performance of this select group over the next five years. The result of one such experiment is

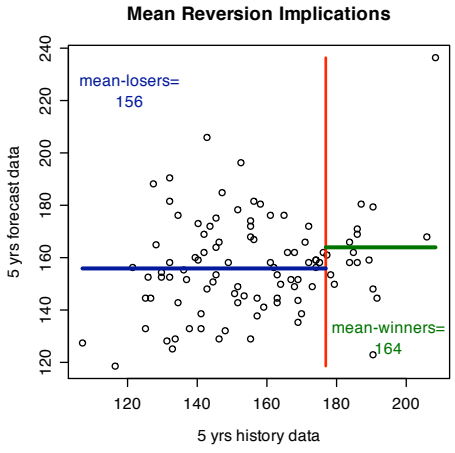


Fig 6: This graph shows that the use of past performance to select fund managers is not a useful strategy. Even if the managers "quality" is unchanged, their quality is swamped by randomness.

While it is difficult to pick a winner among fund managers, the advantage of diversification represented by a mutual fund has value, especially for a small investment. Companies can go bankrupt, but funds seldom do. Another way to demonstrate this aspect is to consider the following "risky" investment. A company offers you the opportunity of investing in their start-up for one year, and you have determined that for each dollar invested the return in one year will be as follows:

Return to $1	Probability
0.00	.25
.50	.25
1.00	.25
4.00	.25

The investment looks risky in the sense that there is only a 25% chance of profit, and a 50% chance of loss. However, what if a portfolio of 25 such opportunities (in different markets, say) were invested in? The result can be simulated and one such typical result is shown in Fig. 7. Suppose we put $4.00 into each of the 25 companies; how risky is that?

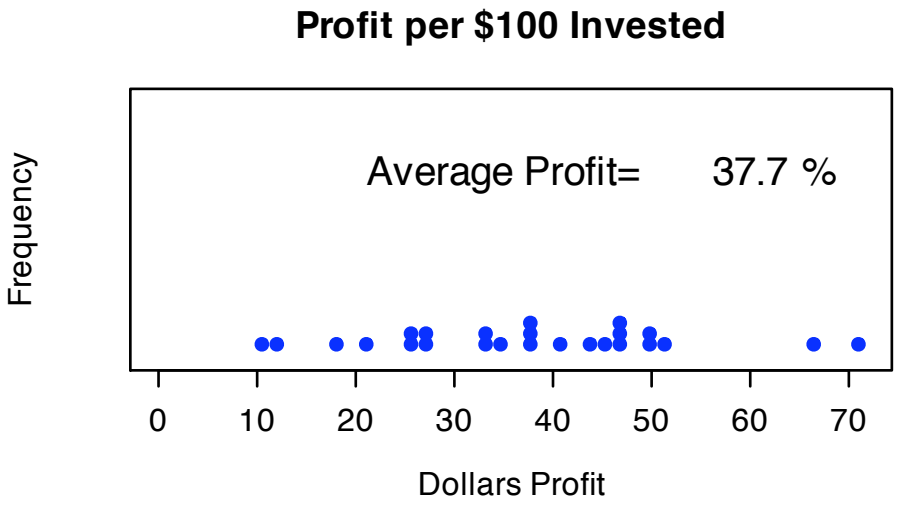


Fig. 7 This scenario shows that a portfolio of risky companies can be profitable.

Obviously, the risk of loss is very small, even if every company in the portfolio is "risky".

Now this demonstration depends on the independence of the 25 companies, and this is only approximately feasible in a real life investment. And also, the demonstation is really just a rework of the square root law, and so it is not so surprising. But it does remind one that investments with a high probability of failure are not necessarily a negative contributor to a portfolio.

In this section, two scenarios have been described for uses of simulation models. In the mutual fund scenario, the property portrayed by the simulation output is that past performance of a mutual fund does not help to identify the future profitability of the fund – the reason was that, although the quality of the manager will affect that manager's prospects, this quality is not identifiable from even 5 years of data, since the inherent

variation in the equity market swamps the effect of manager quality over such time spans. In the second scenario, random variation plays a large role in the outcome of particular investments, but a portfolio of at least partially independent investments can be reliably profitable even when the portfolio consists only of risky investments.

Part III: The advantage of size for simple businesses

It is trite to draw attention to the ability of large companies to squeeze out smaller companies that pose a competitive threat. But there is one industry that allows this ability to be quantified. The insurance industry is very much dependent on "random" variation, where "random" in this case means "unpredictable". The success of the insurance industry confirms that the "randomness" in simulation models does actually capture the real world business model.

Our model in this part is based on a simplified casualty insurance situation. We consider an insurance policy that charges $1460 per annum as a premium for auto insurance. We make the further specification that the cost of a claim is $6000, and that claims occur at the rate of .2 per year, with a Poisson occurrence process determining claims per policy. Although the average claim from this policy is $1200, so that the insurance company will tend to have a gross profit, the randomness of experience can vary this profitability. If we assume the company only has 100 such policies in force, we can simulate many years experience for this company. See Fig. 8.

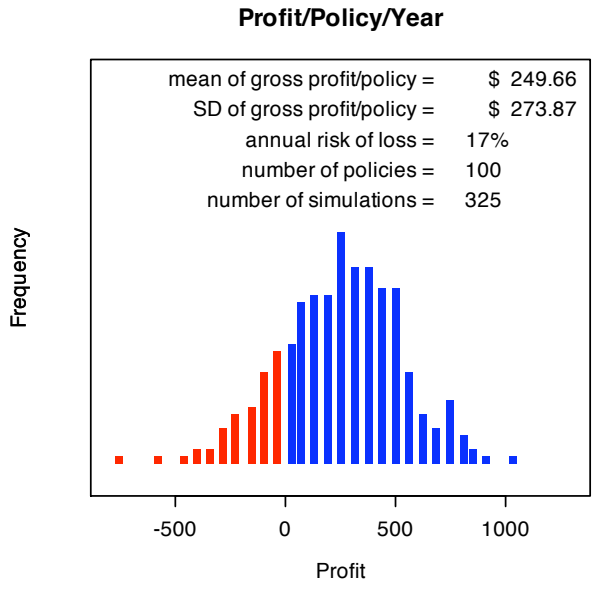
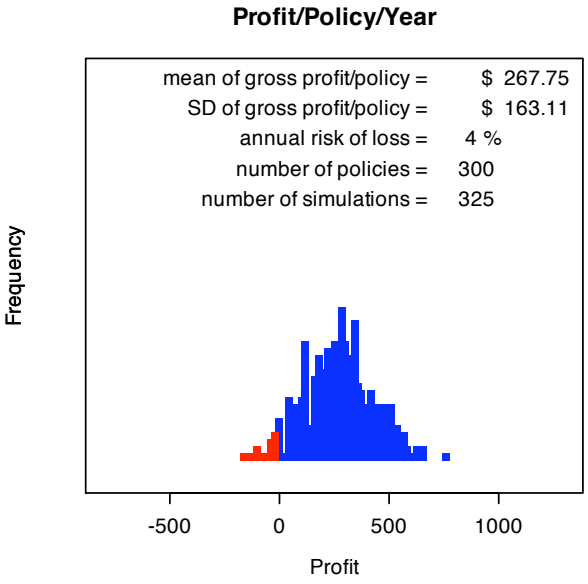


Fig. 8: In 325 years of simulation, the operation lost money in 17% of the years. This small company unit (with 100 policies) needed to be larger to be confident of profitability.

Fig. 9 This larger unit with 300 policies has a more promising prospect of profit, with only 4 % chance of gross loss.



In other words, the larger company unit will likely remain profitable longer than the smaller unit. See Fig 9. Note that a very large unit can reduce its premium to a point where smaller companies will be forced into bankruptcy. See Fig 10.

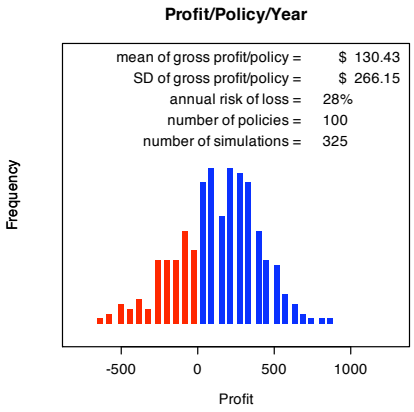
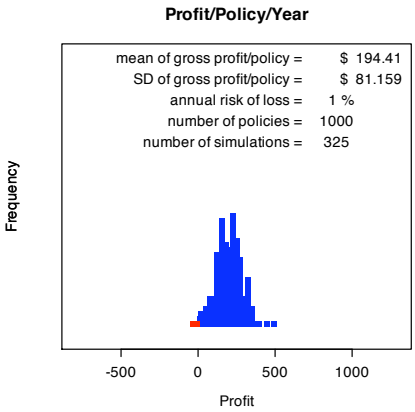


Fig. 10 100 vs 1000 policies, reduced premium,



This is another demonstration of the effect of the square root law in a practical setting. While the statistical theory is absolutely basic, the simulation program that uses it provides a way to explore its effect in an important commercial setting.

Part IV: A simulation model of health status & hospital utilization

A project initiated with Leonard Maclean (Dalhousie University) in 1998 is finally bearing fruit. The expertise of Andrej Blejec (University of Ljubljana) was required to program the health status model in R in 2004. Then I had to find an appropriate place to present the material, and it seems that Dalhousie Business school was a good choice.

The original problem was to find a way to manage the supply-demand relationship of hospital beds in Nova Scotia. Our model is a contribution to that problem. Our approach is to construct a simulation model for the lifetime health status of individuals, to devise triggers for admission and discharge from hospital, and to aggregate these individual case histories over a large population. We then calibrate the model to match known data about the lifetime duration distribution, the length of stay distribution of hospitalizations, the frequency of hospitalizations for individuals, and the distribution of hospitalizations over a lifetime. One use of the model is to explore the impact of varying the beds capacity. Another is to modify the admission and discharge levels.

Figure 11 shows two typical case histories generated by the model.

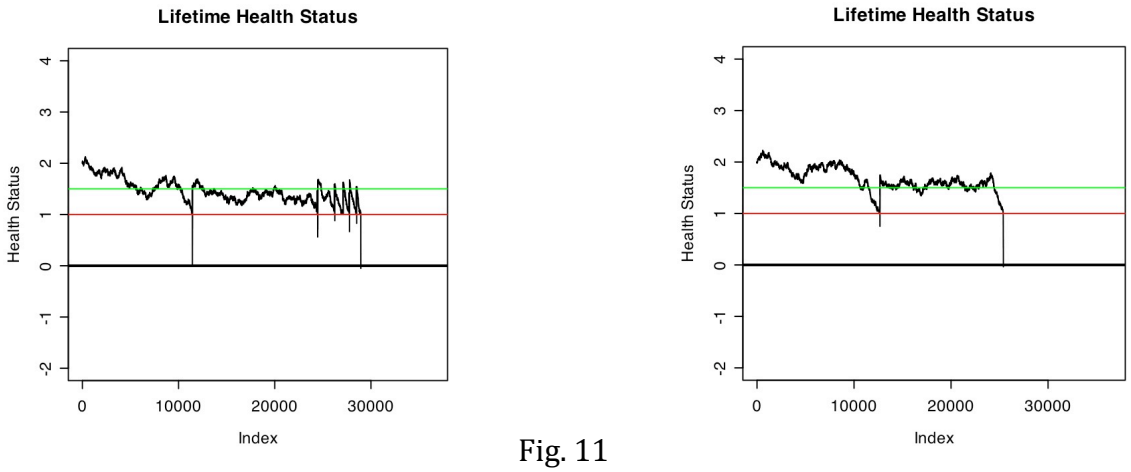


Fig. 11

The green line represents the health status associated with patient discharge, the red line with patient admission, and the black line with death. In the fiction of the model, the first patient had a close call at age 30 (appendicitis, perhaps?) but recovered until some fatal problems occurred during ages 66-79. The second patient had a minor scrape at about age 30 but had a sudden fatal catastrophy at age 69. The simulation is produced on a daily basis for each individual life.

The process allows for a downward drift in health status that is realistic in view of the age-st-death distribution we observe. The daily rate of "cure" while in hospital is usually quite rapid but has a large variability, and can be negative. See Fig 12. The basic model is a random walk with two phases. In the out-of-hospital phase the walk has a slight downward drift. The in-hospital phase has a strong upward drift, but both phases have

different variabilities, and the drift is no guarantee of the direction of change for a particular day.

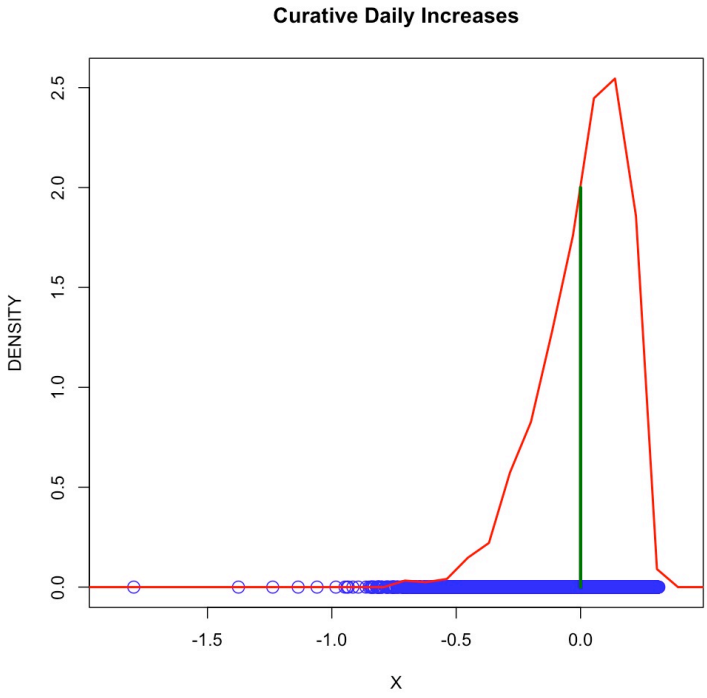


Fig 12
Distribution of Daily changes in Health Status while in the Hospital. Mean is about +0.02, and SD is about 0.2.

The simulation of each of these health status lives takes approximately one second on a modest laptop. To generate enough lives to enable estimation of a steady state on one day, 10000 lives were generated. The aggregate effect of 10000 lives allows us to see what would happen on a random day. To avoid dynamic effects of start-up, we assume the birth process is stationary so that the generated lives can be selected at random times. The effect is to simulate the cross-sectional experience in a population that had been stable for a long time.

The realism of the model can be tested in several ways. For example, the age-at-death distribution of the population, and the length of stay distribution for hospitilizations, can be shown graphically as in Fig 13. Although it is for one simulation of 10,000 lives, other simulations show a very similar pattern.

A median age at death of about 79 years, and a median length-of-stay of 10 days, seems reasonable.

Another benchmark is the number of hospital beds required on a random day. In two simulations of the population of 10,000, the number of beds required was 26 and 31. A rate of .3 per 1000 needs to be compared to the actual bed occupancy rate to determine if this is close enough.

Yet another benchmark is the ages of the individuals that happen to be in hospital on a particular day. Fig 14 shows an example of this.

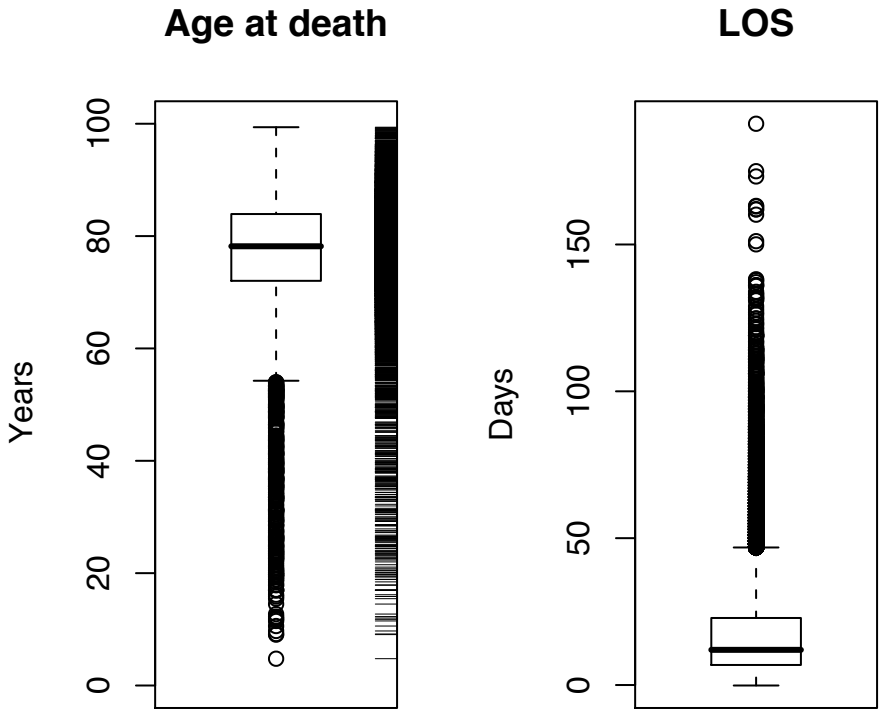


Fig.13

The age at death distribution for the 10,000 simulated lives. And the length of stay distribution for all stays experienced by this simulated population.

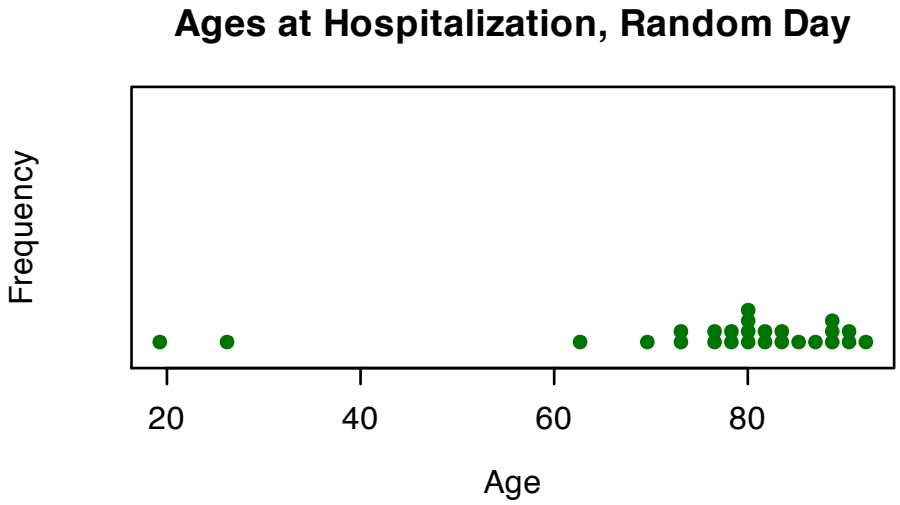


Fig 14.
Age at hospitalization in a cross-sectional sample

As mentioned in the introduction to this part, the model has uses such as determining the ability of a given number of beds to handle the demand, and the effect on bed requirements of changing the admission and discharge policies.

Summary:

R is an open source program for probability and statistics. Its strength is in the ease of construction of non-standard applications, and a frutiful area seems to be in simulations. Simulations can used for analysis or for demonstrations. The four financial applications discussed here have been offered to support the idea that data analysis is sometimes best done using simulation and graphics, and that R is a suitable language to use for the process. Note that R programs can be used with zero expenditure by anyone with internet access, and moreover the use of the programs does not require any special programming knowldege. Of course, there is a learning curve for the programming itself.

References:

Consigli,G., MacLean,C, Zhao,Y. and Ziemba,W.T. (2009) "The Bond-Stock Yield Differential and a Risk Indicator in Financial Markets", Journal of Risk 11(3), to appear.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Stigler, S. (2008) Key Statistical Ideas Celebrate Birthdays. Harvard University Gazette, October 2, 2008. Harvard News Office.

KLW 090215