Spreadsheets in Education (eJSiE)

Volume 2 | Issue 1 Article 5

November 2005

Using an Excel Extension for Selecting the Probability Distribution of Empirical Data

Abbas Heiat
Montana State University, aheiat@msubillings.edu

Using an Excel Extension for Selecting the Probability Distribution of Empirical Data

Abstract
Teaching the steps required for determining the probability distributions of uncertain variables using empirical data is an important part of quantitative and decision analysis courses in business and economics. This paper introduces the concept of distribution fitting of empirical data through and example using an Excel add-in tool.

Keywords
Fitting Probability Distributions, Goodness-Of-Fit Tests, Crystal Ball Excel Add-in

Using an Excel Extension for Selecting the Probability Distribution of Empirical Data

Abbas Heiat
Information Systems
Montana State University
aheiat@msubillings.edu

Abstract

Teaching the steps required for determining the probability distributions of uncertain variables using empirical data is an important part of quantitative and decision analysis courses in business and economics. This paper introduces the concept of distribution fitting of empirical data through and example using an Excel add-in tool.

Keywords: Fitting Probability Distributions, Goodness-Of-Fit Tests, Crystal Ball Excel Add-in.

1. Introduction

Exploring and modeling the distribution of an observed data sample is a key step in many applications of statistics and Monte Carlo simulations. However, very few quantitative analysis or operations management textbooks cover the steps required for determining the probability distributions of uncertain variables using empirical data available for the application. Furthermore, a survey conducted by Albritton, McMullen, and Gardiner indicates that coverage of probability distributions and simulation in undergraduate and MBA business classes is introductory, [1][6].

In this paper I will focus on two steps of using empirical data to model probability distributions of uncertain variables: (1) creating a histogram of the observed data and (2) using goodness of fit tests to determine the most appropriate distribution. Student Editions of Crystal Ball and @Risk Monte Carlo simulation software are available to be used with many quantitative and decision analysis textbooks. While both soft ware packages have distribution fitting capabilities, in this paper I use Crystal Ball's Batch Fit tool to demonstrate a real world business application.

2. Educational Context

I teach an advanced undergraduate decision support systems class (MIS 492). This class covers a range of decision analysis and statistical topics, including simulation. Most students are required to take an introductory business statistics course prior to this course, so they have had some exposure to statistical topics, but few students have any academic experience with simulation. Due to the typically wide variety of student skills, teaching simulation presents a challenge for both students and instructors. The course begins with a discussion of basic concepts in simulation and demonstration of several examples. Once the fundamentals are covered, using a problem based approach; I use a

real world case to provide practice in the skills that students will need for decision making and problem solving: the ability to collect data, build models, and conduct analyses in a timely and cost effective manner. Once students have a grasp of this ability, they should apply the learned skills to their class project. The problems to be analyzed in students' projects can be addressed via the features of the spreadsheet simulation software specifically via the Crystal Ball's distribution fitting feature. There is almost always a reaction of pleasant surprise from students when they learn how to model uncertain data. With more extensions of spreadsheet simulation presented here, even more "real-world" functionality can be added to the business analysis experience of students.

3. Simulation of Fixed Charge Coverage Ratio

I use the following real world example in my advanced undergraduate decision support systems class to teach determining probability distributions for simulation analysis. Banking analysts must judge the credit worthiness of new or existing clients based largely on a financial concept known as the Fixed Charge Coverage Ratio (FCCR). The ability to accurately predict FCCR depends on the ability to accurately predict underlying financial measures such as Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA), Cash Taxes Paid, Cash Dividends Paid, and Unfinanced Capital Expenditures. In general, the prediction of FCCR was completed using historical averages. However, Monte Carlo simulation can provide a method for more accurately simulating and predicting the FCCR of a particular type of loan. The amount of Principal that will be paid within the next year by the firm is considered the Mandatory Debt Retirement. Following is the equation for FCCR:

$$FCCR = \frac{EBITDA - Cash\,Taxes\,Paid - Unfinanced\,Capital\,Expenditure}{Interest\,Expense + Mandatory\,Debt\,Retirement}$$

In credit analysis, the use of the FCCR is extremely important. First, it attempts to place all potential loans on a common ground for comparison. Each element of the calculation of the Fixed Charge Coverage Ratio is standardized and defined in fairly specific terms. Second, the FCCR is relatively easy to calculate. The FCCR is prominent on the decision documentation for loans and is often the first thing a credit administrator will look for when making a loan decision. As the FCCR increases, it is assumed that the ability to service debt increases. For instance, a FCCR of 2:1 or 200% would indicate that the Borrower has twice as much earnings to service debt than debt service. If the Borrower had a FCCR of 0.5:1, it is assumed that only half of the debt service could be met with earnings. While not codified in the credit policy, the general rule for a "good" FCCR is 1.25:1 or 125%. This means that in order for a credit to be deemed capable of sufficient debt service, the Borrower must have earnings greater than 125% of proposed debt service.

4. Sample of Empirical Data

The data used for the simulation was gathered from the Small Business Banking Database for the State of Montana. The database contains data for approximately 2000

loans. Of these 2000 loans, 30 were randomly chosen and four financial measures were gathered for each of the 30 loans.

There are many reasons a business analyst would like to use a representative sample of existing data. They include: costs, speeding up the data collection, improving effectiveness, and reducing bias, [3]. Students, analyzing real world cases, often realize that examining and using every piece of information would be far too costly and they need to learn and use sampling. In our case, calculation of FCCR for 2000 records would be a time consuming process and it is impossible to do it during the class time allocated for this exercise. Using a random number generator, 30 numbers from the 2000 were randomly chosen. Next, the financial measures Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA), Cash Taxes Paid (CTP), Cash Dividends Paid (CDP), and Unfinanced Capital Expenditures (UCEXP) on each of 30 loans were calculated. The sample data for the four uncertain variables are listed in Table 1.

Table 1: Sample Data for Uncertain Variables

EBITDA	CTP	CDP	UCEXP
-128620	0	0	-20379
-108598	0	0	-9970
-82202	0	0	-9825
-69800	0	0	-2328
-17889	0	0	499
-8902	0	0	849
-1735	0	0	994
25726	0	0	1415
30072	0	0	3030
83830	0	2601	4079
84344	0	4612	4294
103844	0	5109	4751
108050	256	5377	5719
128981	12395	6057	8127
132344	26254	6318	8212
157204	28045	6840	9496
161915	32972	6887	9749
165052	35690	7536	9771
211702	41392	7584	11687
232850	44105	7755	15101
234134	50883	8042	15409
254408	54157	8334	15577
286374	56254	8451	20258
298405	60284	8559	21045
302525	68769	8904	22224
316471	69730	10832	23890
321490	89753	11854	24744
349475	104584	13531	25368
474955	120876	15593	26084

5. Visualizing the Observed Data

In Monte Carlo simulation and many other business statistical analyses, business analyst must make assumptions about the uncertainty of one or more inputs. We characterize this uncertainty by specifying a probability distribution for uncertain inputs.
To select an appropriate distribution, we might start by examining a histogram of the data to see if it is compatible with the shape of any distribution, [2]. Using the Chart Wizard in Excel, the data for Cash Dividends Paid (CDP) is depicted in Figure 1. The Normal distribution is symmetric with a peak in the middle. Exponential and Lognormal distributions are positively skewed. CDP data does not seem to fit any of these distributions. However, various forms of other distributions could be used to find a fit for our data. A visual approach, as demonstrated in the case of CDP data, is not always easy, accurate, or valid to apply especially if sample size is small.

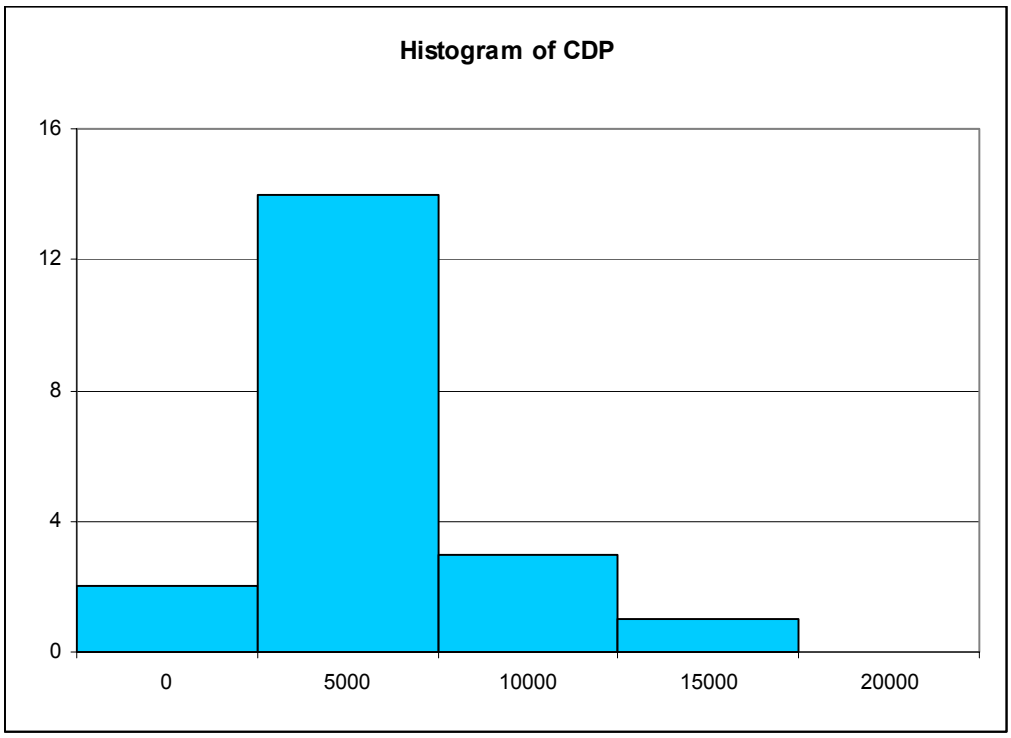

Figure 1: Distribution of Cash Dividends

6. Goodness of Fitness Tests

Crystal Ball provides facilities for fitting and plotting continuous distributions. When you fit sample data to a distribution, Crystal Ball provides a series of goodness-of-fit tests and *p*-values based on the empirical distribution function (EDF). The EDF tests, including the Kolmogorov-Smirnov, Chi-Square, and Anderson-Darling are based on various measures of discrepancy between the empirical distribution function and the cumulative distribution function based on a specified distribution, [2], [4] and [5].

Using Crystal Ball, to fit data to a distribution, students should select **Define Assumptions** from **Cell** menu. The **Distribution Gallery** is displayed. Next, students click on **Fit** button which brings up the **Fit Distribution** window. In the Fit Distribution window the input range should be defined. The input range is the column in Excel that contains the sample data for the intended variable. Once the input range is entered, students click **Next** which displays the second Fit Distribution window. Crystal Ball displays the results for all three tests; Kolmogorov-Smirnov, Chi-Square, and Anderson-Darling. However, students may choose the most appropriate test for ranking the results. **The Comparison Chart** and test results are displayed for several distributions ranked by the ranking order selected in the previous step. Students could navigate through results for various distributions by using **Next Distribution** button. Figure 2 shows the first fitted distribution for Cash Dividends Paid (CDP) selecting Anderson-Darling as the ranking method.

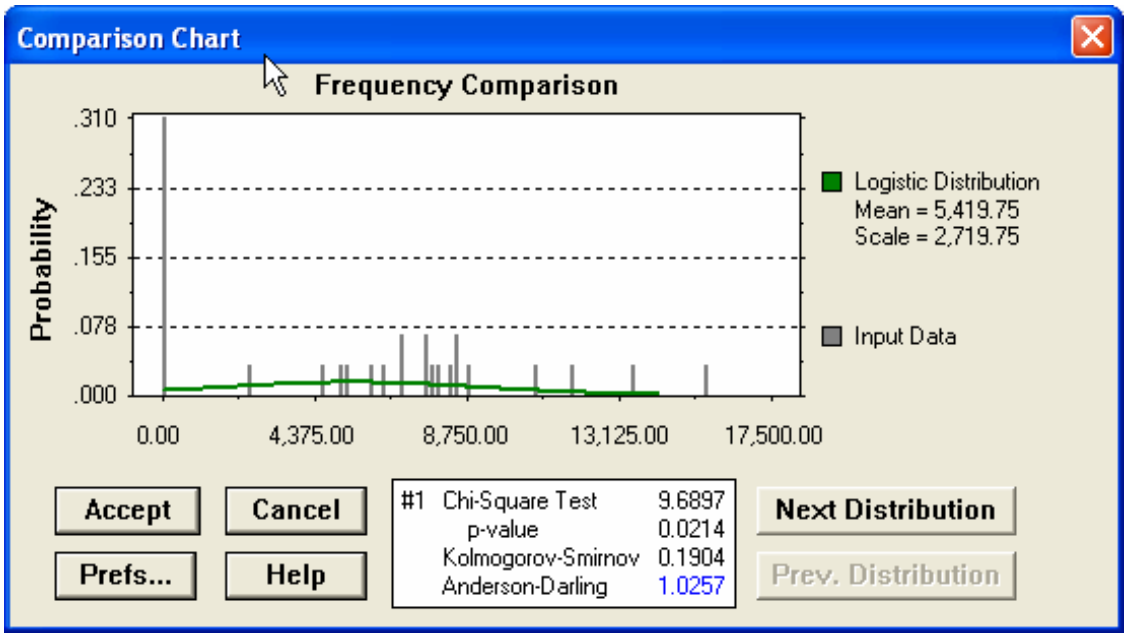

Figure 2: Crystal Ball Comparison Chart Showing Best-Fitting Distribution

Based on Anderson-Darling test, the Logistic Distribution is the best distribution that fits our sample data for Cash Dividends Paid (CDP).

The **Chi-square test** is used to test if a sample of data came from a population with a specific distribution. The chi-square goodness-of-fit test is applied to binned data i.e., data put into classes. As a general rule, we should have at least 50 observations to perform the test. If the calculated value of chi-square test is less than the critical value with given probability and degree of freedom, we fail to reject the hypothesis that the data come from the fitted distribution. For example, in Figure 2 the value of chi-square test is 9.6897 which is less than the critical value in the table (14.256) with 0.99 given probability and 29 degree of freedom. Therefore, the Logistic Distribution seems to be a good fit.

For small samples, the **Kolmogorov-Smirnov (K-S)** test provides a better choice. Despite this advantage, the K-S test only applies to continuous distributions and it tends to be more sensitive near the center of the distribution than at the tails. If the calculated value of Kolmogorov-Smirnov test is greater than the critical value with given probability and degree of freedom, we fail to reject the hypothesis that the data come from the fitted distribution. For example, in Figure 2 the value of Kolmogorov-Smirnov is 0.1904 which is less than the critical value of 0.295 with given probability of 0.99. In this case, the K-S test indicates not a good fit.

Due to limitations of Chi-Square and K-S tests, many analysts prefer to use the **Anderson-Darling goodness-of-fit test**. However, the Anderson-Darling test is only available for a few specific distributions. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test. A computed value less than 1.5 generally indicates a good fit. In our example, the computed value of Anderson-Darling is 1.0275 which means that Logistic Distribution is a good fit for our sample data, [2]. Once the distributions for uncertain variables are determined, we may build the model in Excel and use Define Assumption menu to define uncertain variables or assumptions as they are called in Crystal Ball. Next step is to run simulation and conduct probability analysis for FCCR predicted value.

7. Conclusion

Exploring and modeling the distribution of an observed data sample is a key step in simulation and many applications of quantitative and decision analysis in business and economics. Excel add-in software provide an easy distribution fitting methods of empirical data. This add-in is integrated software that includes optimization and simulation tools. They are easy to use and provide for exporting transparently the fitted distributions into Excel.

References

[1] Albritton, M.D., McMullen, P.R., and Gardiner, L.R., "OR/MS content and Visibility in AACSB-accredited US Business Programs", Interfaces, 2003.

[2] Evans, D.R., and Olson, D.L., Introduction to Simulation and Risk Analysis, Upper Saddle River, Prentice Hall, 2002.

[3] Kendall, Kenneth E. and Julie E. Kendall, Systems Analysis and Design, Sixth Edition, Upper Saddle River, Prentice Hall, 2005.

[4] Liu, Jun S. Monte Carlo Strategies in Scientific Computing, New York: Springer, 2001.

[5] Mascialino, B. and Pia, M. G., A Toolkit for Statistical Comparison of Data Distributions, http://www.ge.infn.it/geant4/papers/2005/mc2005/statistics.pdf.

[6] McMullen, Patrick R., "Using Correlation Matrices and Optimization to Add Practical Functionality to Spreadsheet Simulation for MBA-level Quantitative Analysis Course", Decision Sciences Journal of Innovative Education, January 2005.