BASEBALL AND STATISTICS: AN EXPLORATION USING EXCEL IN A FIRST-SEMESTER STATISTICS COURSE

Raimundo Kovac, Rebecca Sparks, Christopher Teixeira
Rhode Island College
Department of Mathematics and Computer Science
600 Mount Pleasant Avenue
Providence, RI 02908
rkovac@ric.edu, rsparks@ric.edu, cteixeira@ric.edu

1. Introduction

With all students now having access to Excel and its convenient statistical functions, even the most introductory level students can work with large amounts of data and interesting examples that used to be awkward for classroom and homework use. Because of this old "awkward" problem, classroom examples in statistics became contrived and dry. We describe two classroom projects that use Excel to handle two topics with which first-semester statistics students struggle to obtain deep understanding: the Central Limit Theorem and Multiple Regression. The projects can certainly be done independent of Microsoft Excel; other software packages do a fine job. We chose Excel because of its easy availability, and we do prefer using a computer over the graphing calculator due to the huge data sets, and, the high quality graphs that Excel (or other packages) provides over the typical graphing calculator.

For our demonstrations, we use statistics from the 2006 Season of Major League Baseball. Baseball and statistics is a marriage that has been celebrated over generations, and we exploit this in our first demonstration by using a pitcher's earned run average (ERA) to demonstrate the power of the Central Limit Theorem. The beauty of such an example is that we have access to the underlying population – an unusual situation in statistics – and so can specifically illustrate the changes that occur in the distribution of the sample mean as the sample size changes. This is something that can be done efficiently using the current software; not long ago it would have been completely impractical.

2. The Central Limit Theorem

To illustrate the Central Limit Theorem (CLT), we use starting pitchers who accumulated at least 162 innings pitched. (This allows for, on average, one inning pitched per game.) The population parameters are as follows:

- N=77
- Mean, $\mu = 4.275$
- Standard Deviation, $\sigma = 0.7405$
- Minimum ERA = 2.77
- Maximum ERA = 6.36

Although the Central Limit Theorem does not require it, the data set described above is not a normal distribution. This does however, make for a more interesting demonstration, particularly for most skeptical of students.

The Central Limit Theorem can be stated as follows: If the random variable x has any distribution with mean μ and standard deviation σ, and samples of size n are randomly selected from the population, then:

1. The distribution of all possible sample means, $\bar{x}$,will approach a normal distribution as the sample size increases.

2. The mean of the sample means will be μ.

3. The standard deviation of the sample means will be $\sigma/\sqrt{n}$. If however, the population is finite (size N), and n is greater than 5% of N, then the sampling distribution has standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

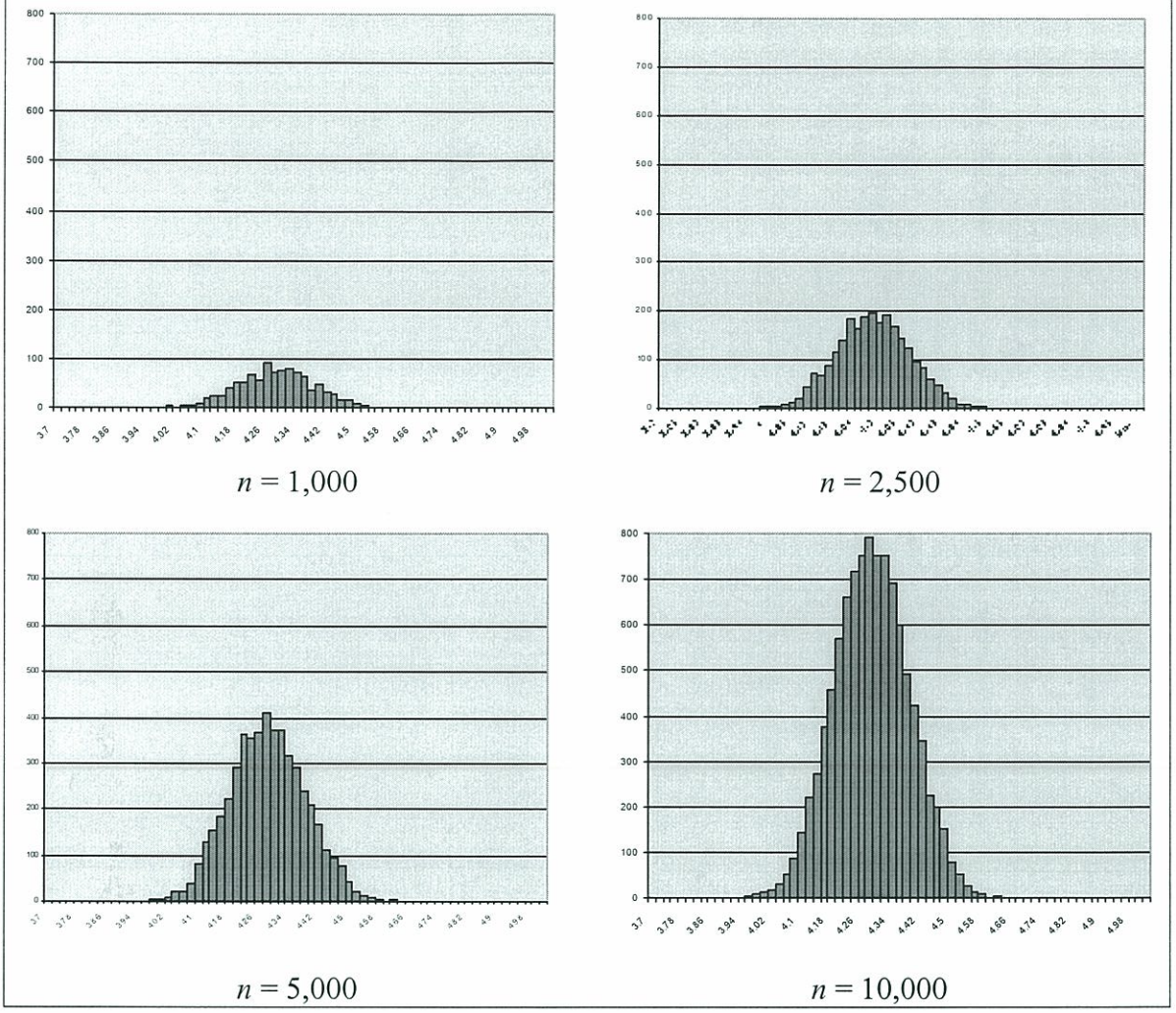
For the purposes of demonstration in class we use samples of size 2, 4, 6, 16, 32, and 64. For each sample size, histograms are created for increasing numbers of samples. In addition, we calculate the mean and standard deviation for each frequency distribution. Table 1 is an example of the frequency distributions for samples of size 32. The student can quickly draw the following conclusions.

1. As sample size increases we get a better approximation for the normal distribution. (The frequency distributions for samples of size 2 look less "bell-shaped" than the frequency distributions for samples of size 32.)

2. The mean and standard deviation of the sample means actually is close to that predicted from the CLT. (For example, with our samples of size 32, in the case of 10,000 samples, CLT predicts that the standard deviation should be 0.1007, and the mean of the sample means should be 4.275. The actual standard deviation is 0.0995, and the actual mean of sample means is 4.276.)

3. CLT tells us that the standard deviation should decrease as sample size increases. Graphically, they may now see this as thinner bell curves. Our students were most interested in seeing this behavior, and as a result, started referring to this behavior as the "incredible shrinking sigma".

Table 1: Frequency Distributions, Samples of Size 32



$n = 1{,}000$ $n = 2{,}500$

$n = 5{,}000$ $n = 10{,}000$

3. Multiple Regression

Next, we use MLB statistics to demonstrate the use of multiple regression by trying to predict the Cy Young Award winner. This award honors the "most outstanding pitcher" on each league. The winner is selected by a preference ballot using a 5-3-1 Borda count. The votes are cast by members of the Baseball Writers Association of America, with each MLB city getting two voters that may vary from year to year. Voters may use their own criteria to decide their preference order, from the wealth of publicly available statistics to their own subjective opinions of each pitcher's worth.

In this activity, we want to find out if the winners of the award can be predicted based on the available statistics for each player. We emphasize that we do not want to speak to who "deserves" to win, but who the writers will choose with their voting. Is it possible to take the season statistics for each pitcher in the league and, using multiple regression,

come up with a formula that will predict the winner? Could we go as far as predicting the first-second-third finishing order and their total points?

Our model included only starting pitchers. Although in recent years, the "closer" position has become a bit of a glamour position for pitchers, historically relief pitchers were never considered to be among the league's elite. As a result, relief pitchers rarely appear on the ballots of voters, and, only a handful of closers have ever won the award. Moreover, their pitching statistics are so different in comparison to the many innings of an ace starting pitcher, we decided to eliminate all relievers from the model in order to give ourselves a fighting chance. We decided that the most relevant statistics to consider were pitcher's wins (W), losses (L), earned run average (E), and strikeouts (K). We suspected that team winning proportion (T) would be relevant too – pitchers in winning teams get more visibility – and included this statistic as well. We used the data from 1993 to 2005, for both the American League and the National League. Thus, we set up an Excel spreadsheet with the first, second, and third place winners for each year in rows, with a column for each of the above mentioned statistics, and a column for the points P earned by each pitcher in the voting. Then we used Excel's multiple regression tool to find five constants m_1, m_2, m_3, m_4, and m_5 so that

$$P = m_1W + m_2L + m_3E + m_4K + m_5T\,.$$

The results we obtained were as follows:

$$m_1 = 6.7329,\ \ m_2 = -1.3235,\ \ m_3 = -27.201,\ \ m_4 = 0.994,\ \ m_5 = 4.1483$$

At first, the students were curious about why m_2 and m_3 were negative. Remember, m_2 is associated with a pitcher's losses in one season. The more losses a pitcher accumulates, the worse it should be for him in obtaining points in the balloting process. Also, m_3 is associated with a pitchers earned run average. The larger a pitcher's ERA, the worse it should be for him in obtaining points from a ballot.

Students were also interested in trying to decide which statistics appeared more important to the voters based on the results of the weights, but we must point out that with this exercise, it is impossible. In order to attempt to draw a conclusion in that direction, we would have to go through a normalization process that we felt made the problem more complicated than necessary at this point.

The correlation coefficient was a very good 0.9009, but, how good was the regression equation when applied to the data set? As it turned out, the prediction of the actual points were not very accurate. However, the equation correctly predicted the first, second, and third place finishers in 22 out of the 26 data sets.

When this project was done in class, the winners for the 2006 season were about to be announced, so we used our equation with the newly available statistics to try to predict

who they would be. Our predictions and the actual outcomes are summarized in the table below.

Table 2: 2007 Cy Young Award Predicted and Actual Results		
	Predicted Points	Actual Points
American League		
Santana	96.070	140
Wang	39.003	51
Halladay	42.987	48
National League		
Webb	50.355	103
Hoffman	45.279	77
Carpenter	44.640	63

The results are consistent with those for our starting data set: both winners were predicted correctly, and in the National League we even predicted the correct finishing order. On the other hand, the total number of points predicted was in general quite inaccurate.

References

Sparks, Rebecca L. and David L. Abrahamson. *A Mathematical Model to Predict Award Winners*. Math Horizons 12(3): 5-9, 2005.

MVP and Cy Young Award Winners. Accessed October 2006; available from http://www.baseball-reference.com/awards/mvp_cya.shtml.

Major League Baseball Statistics Index. Accessed October 2006; available from http://sports.espn.go.com/mlb/statistics.