

## **Fantasy Baseball with a Statistical Twist**

**Dr. Lori Koban**

ASSISTANT PROFESSOR

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

WESTERN CAROLINA UNIVERSITY

EMAIL: LKOBAN@EMAIL.WCU.EDU

PHONE: (828)227-2484

FAX: (828)227-7240

ADDRESS: WCU, DEPT. OF MATH & C.S., CULLOWHEE, NC 28723

**Dr. Erin McNelis**

ASSISTANT PROFESSOR

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

WESTERN CAROLINA UNIVERSITY

EMAIL: EMCNELIS@EMAIL.WCU.EDU

PHONE: (828)227-3947

FAX: (828)227-7240

ADDRESS: WCU, DEPT. OF MATH & C.S., CULLOWHEE, NC 28723

MAY 3, 2006

### **Abstract**

We introduce a Fantasy Baseball Project that is appropriate for entry-level Statistics classes. The project's first objective is to use the statistical methods learned in such a course to analyze baseball data. Though each student can successfully complete the project, it has a competitive nature as well. The project's second objective is to choose the "best" 10-person fantasy team in the class.

## Fantasy Baseball with a Statistical Twist

### 1. INTRODUCTION

Fantasy baseball, a game invented in 1980, allows baseball fans to become managers of pretend baseball teams[1]. In most fantasy baseball leagues, players choose teams that they believe will do well in five offensive categories (batting average, home runs, runs batted in, stolen bases, and runs scored) and in five pitching categories. We bring fantasy baseball into entry-level Statistics classes. Each student drafts a team based on nine offensive categories, most of which are statistical twists on the five categories above. The primary goal of this project is to apply the material in a one-semester, introductory, non-calculus-based, college course in Statistics. Since this is the type of course that AP Statistics courses are designed to emulate, this project is appropriate for AP Statistics classes as well. Indeed, this project incorporates exploratory analysis, planning and conducting a study, probability, and statistical inference, the four major themes of an AP Statistics class [2].

The fantasy baseball project serves as a common thread uniting a variety of topics. Ideally, students work on it throughout the course. An interesting feature of this project is its competitive component. Though all students can complete the project successfully, a class winner can be named. Baseball fans might find this project particularly appealing, but students need not have an understanding of baseball in order to successfully complete (and win) the project. The winner is the person who best analyzes the data. So have fun, and play ball in your Statistics classroom.

### 2. THE FANTASY BASEBALL PROJECT

**Project Description:** From a pool of 25 Major League Baseball (MLB) players, you will choose a 10-player team. Your choice of team will be based on the statistical analysis of several categories. You can find the pool of 25 players in Appendix A.<sup>1</sup> These 25 players are a simple random sample of all players in MLB with at least 350 at-bats during the 2005 season. The lower bound of 350 ensures that our sample consists of players who were offensively active during a majority of the season.

This project consists of 10 parts. Parts 1-9 require a statistical analysis of each of the BEST TEAM CHARACTERISTICS given below. Part 10 is where you choose your fantasy team and compute the values

---

<sup>1</sup>To choose your own sample, go to [sports.espn.go.com/mlb/stats/batting?league=mlb](https://sports.espn.go.com/mlb/stats/batting?league=mlb).

of each characteristic for this team. Although this project is divided into ten parts, it has two main objectives.

**Objective #1:** Use the statistical methods we learn in class to analyze each of the following BEST TEAM CHARACTERISTICS.

**BEST TEAM CHARACTERISTICS**

- (1) Most runs batted in per player
- (2) Fewest strike outs per player
- (3) Most home runs per player
- (4) Highest batting average per player
- (5) Least variation in stolen bases per player
- (6) Strongest linear relationship between number of runs and number of home runs
- (7) Smallest probability that a randomly chosen player has less than 10 stolen bases
- (8) Most narrow 95%-confidence interval for the mean batting average in MLB
- (9) Smallest  $p$ -value for the hypothesis test:  $H_0 : \pi_{\text{highAVG}} = .5$  vs.  $H_A : \pi_{\text{highAVG}} > .5$

**Objective #2:** Be the person with the best fantasy baseball team in your class. Among your classmates' fantasy teams chosen in Part 10, suppose there are  $n$  different values for a given characteristic. If your fantasy team is the best for this characteristic, you will get  $n$  points; if your team is second best, you will get  $n - 1$  points; if your team is the worst, you will get 1 point. You win if you have the highest cumulative score in the class.<sup>2</sup>

**Part 1: Most runs batted in per player**

**Description:** A player's number of runs batted in appears in the column labeled RBI. Two interpretations of the word "most" are: the highest mean of your 10 values and the highest median of your 10 values. For this part we will use the highest mean interpretation.

**Analysis:** Make a graphical display of the RBI data (all 25 points). You can choose from: histogram (with equal or unequal class-interval lengths), stem-and-leaf plot, boxplot, and dotplot. If you choose to make a histogram, include a frequency table. What 10-man team should you choose if "most" means "highest mean?" Is there a single answer to this question? Explain.

<sup>2</sup>For information on programs that have been written to grade this project quickly, please contact the authors.

### Part 2: Fewest strike outs per player

**Description:** The number of times a player struck out is in the column labeled SO. Two interpretations of the word “fewest” are: the smallest mean of your 10 values and the smallest median of your 10 values. For this part we will use the smallest mean interpretation.

**Analysis:** Make a graphical display of the SO data (all 25 points). Your choice here must be different than your choice in Part 1. What 10-man team should you choose if “fewest” means “smallest mean?”

### Part 3: Most home runs per player

**Description:** A player’s number of home runs is in the column labeled HR. For this part we will use the “highest median” interpretation of the word “most.”

**Analysis:** Make a graphical display of the HR data (all 25 points). Your choice here must be different than your choices in Parts 1 and 2. What 10-man team should you choose if “highest” means “highest median?” Is there a single answer to this question? Explain.

### Part 4: Highest batting average per player

**Description:** A player’s batting average is computed by dividing his number of hits by his number of (H/AB). For this part we will use the “highest median” interpretation of the word “highest.”

**Analysis:** This data does not appear in the data set directly, so insert an “AVG” column. Make a graphical display of the batting average data (all 25 points). Your choice here must be different than your choices in Parts 1, 2, and 3. What 10-man team should you choose if “highest” means “highest median?”

### Part 5: Least variation in stolen bases per player

**Description:** The number of bases that a player stole is in the column labeled SB. Three interpretations of the word “variation” are: the range, the interquartile range, and the standard deviation of your 10 values.

**Analysis:** Make a dotplot of the SB data (all 25 points). What 10-man team should you choose if “variation” means “range?” What if “variation” means “interquartile range?” What if “variation” means “standard deviation?” Use your dotplot to help explain why you chose each team.

### Part 6: Strongest linear relationship between number of runs and number of home runs

**Description:** You want a team for which the number of runs (R) and the number of home runs (HR) have the strongest linear relationship.

**Analysis:** Make a scatterplot of R vs. HR. (Place R on the horizontal axis.) What 10-player team appears to have a very strong linear relationship between these variables? Identify your team on the scatterplot. For this team, what is the correlation coefficient?

**Part 7: Smallest probability that a randomly chosen player has less than 10 stolen bases**

**Description:** Randomly choose one player from a 10-player team. Since we know each player's number of stolen bases, we are able to compute the probability that a randomly chosen player has less than 10 stolen bases.

**Analysis:** What 10-player team should you choose to minimize this probability? Is there a unique answer?

**Part 8: Most narrow 95%-confidence interval for the mean batting average in MLB**

**Description:** Each 10-player team is a sample of MLB players and yields a confidence interval for  $\mu_{\text{AVG}}$ , the mean batting average in MLB. (Technical conditions: To make this confidence interval, the sample should be random and the distribution of MLB batting averages should be approximately normal [3].)

**Analysis:** RANDOMLY choose 10 of the 25 players, and use this 10-player team to make a 95%-confidence interval for  $\mu_{\text{AVG}}$ . Explain your procedure for choosing the 10 players randomly. What 10-player team yields the most narrow 95%-confidence interval for  $\mu_{\text{AVG}}$ ? Explain why you chose this team. Do you think that the technical conditions for this confidence interval are satisfied? Explain.

**Part 9: Smallest  $p$ -value for the hypothesis test  $H_0 : \pi_{\text{highAVG}} = .5$  vs.  $H_A : \pi_{\text{highAVG}} > .5$**

**Description:** Define  $\pi_{\text{highAVG}}$  to be the proportion of batting averages in MLB that are greater than .270. Each 10-player team can be considered a sample of MLB players and can be used to conduct the hypothesis test stated above. (Technical conditions: To conduct this hypothesis test, the sample should be random and large enough [3].)

**Analysis:** Using the 10-player team that you randomly chose in Part 8, conduct this hypothesis test. What is the  $p$ -value? Should you reject the null hypothesis? What 10-player team do you think would yield the smallest  $p$ -value for this hypothesis test? Explain why you chose this team. Do you think that the technical conditions for this hypothesis test are satisfied? Explain.

### Part 10: Your 10-Player Fantasy Team

**Description:** It's time to choose your fantasy team. Remember, you want to choose the team that you think will give you the highest cumulative score.

**Your 10-player team:** Give each player's name and what number they are in the original data set. Briefly explain why you chose this team. For your team:

- What is the mean number of runs batted in?
- What is the mean number of strike outs?
- What is the median number of home runs?
- What is the median batting average?
- What is the standard deviation of stolen bases?
- What is the value of the correlation coefficient for runs vs. home runs?
- What is the probability that a randomly chosen player has less than 10 stolen bases?
- What is the width of your 95%-confidence interval for  $\mu_{\text{AVG}}$ ?
- What is the  $p$ -value for the hypothesis test  $H_0 : \pi_{\text{highAVG}} = .5$  vs.  $H_A : \pi_{\text{highAVG}} > .5$ ?

### 3. THE PROJECT WITH THE TI-83

The data set is manageable enough that all of the statistical analysis can be carried out by hand. However, appropriate use of technology not only makes calculations easier, but it also enhances students' visual understanding and intuition about the problems. We have looked at ways of best using four distinct types of technology to aid in the completion of the Fantasy Baseball project: the TI-83, Microsoft Excel®, Fathom™, and MINTAB®. Since the TI-83 is the tool most accessible to students and educators, we will address solving the Fantasy Baseball Project with this graphing calculator. For similar descriptions of how to use the other tools in conjunction with the Fantasy Baseball Project, please contact the authors.

The TI-83 has built-in statistical features ranging from calculating basic statistical measures to making inferences about population parameters using inferential tests. The fantasy baseball project can be tackled using a thorough understanding of statistics along with the simplest statistical features on the TI-83.

Start by inputting each column of the player data into a TI-83 LIST. For clarity, we recommend that the `SetUpEditor` be used to predefine the lists named: NUM, AB, H, R, SB, SO, HR, RBI and AVG (see Figure 3.1(a)). The data is most easily entered into the TI after saving it (minus player names) in a *tab delimited*

text format from a spreadsheet, and importing the data as a matrix via the TI-Graph link. The matrix is converted to lists through the `Matr►list` function, as is shown in Figure 3.1(b).

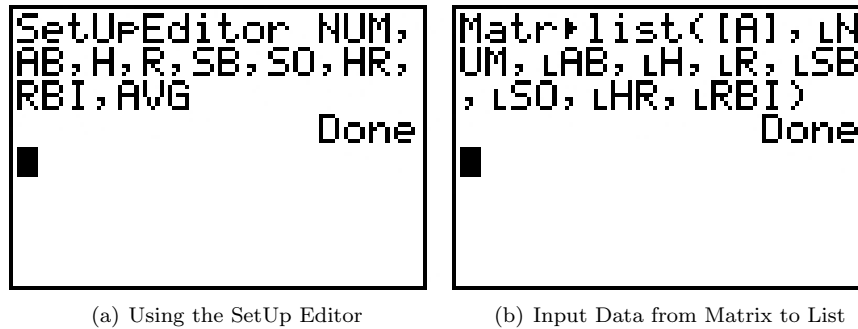


FIGURE 3.1. Getting the Data into the TI-83

Note that the list `AVG` was created with the `SetUpEditor`, but was not assigned values with the `Matr►list` function. This is because batting averages were not included in the original data, but were intended to be created using the at-bats (`AB`) and hits (`H`) data. Simply define the `AVG` list to be `LH/LAB`, as illustrated in Figure 3.2.

HR	RBI	AVG
11	55	-----
31	92	
3	47	
15	83	
7	66	
8	70	
15	68	
AVG = LH / LAB		

FIGURE 3.2. Defining the Batting Average in `AVG` List

Although the fantasy baseball project incorporates topics and techniques from a semester-long introductory statistics course, eight of the first nine parts can be completed using knowledge of the statistical concepts and a simple sorting and graphing of the data. All but Part 6's challenge to find a team with a very strong linear relationship can be approached by first sorting the data by the appropriate column. We focus here on solving three of the first nine parts of the Fantasy Baseball Project with the TI-83. Solution procedures for the entire project can be found in Appendix B.

**3.1. A Team with Least Variation in Stolen Bases.** Part 5 asks us to find the team with the least variation in stolen bases per player, where variation is measured in terms of range, interquartile range, and

standard deviation. Start by sorting all of the data in terms of ascending SB values using the `SortA` command. In order to keep the data aligned, it is important to include all other lists after `LSB`, as is illustrated in Figure 3.3.

```
SortA(LSB, LNUM, L
AB, LH, LR, LSO, LHR
, LRBI, LAVG)
Done
LSB
(0 0 0 1 1 1 1 ...
```

FIGURE 3.3. Sorting Lists by Stolen Bases

The first ten players in the sorted list have a range of 2 stolen bases, which happens to be the smallest possible range for any team of ten players. Figure 3.4 indicates that the second through eleventh players would give the same result, so our answer is not unique!

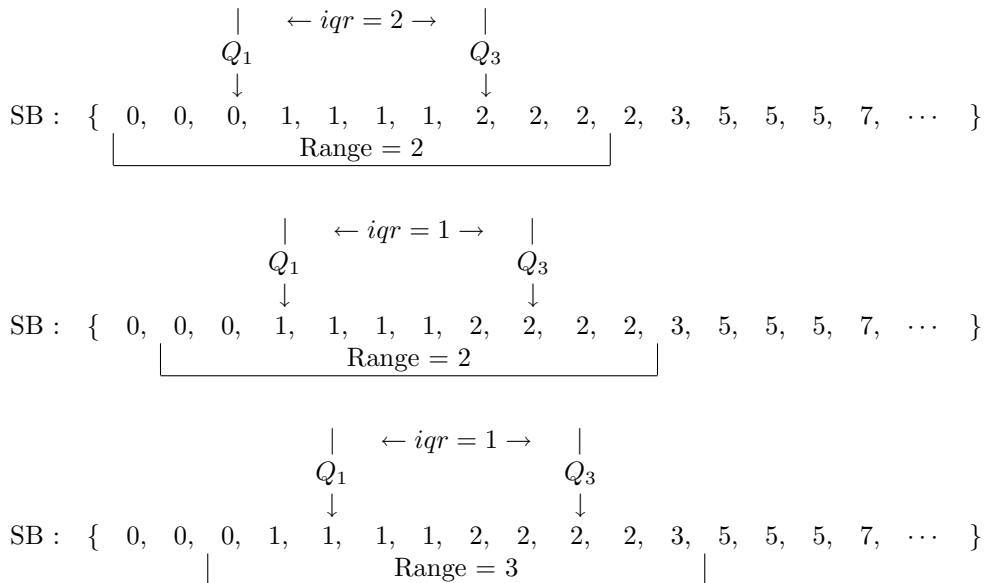


FIGURE 3.4. Comparing Range and Interquartile Range

For the least amount of variation in terms of interquartile range, look at the tightly clustered values from 0 to 3. Keeping in mind that the first and third quartiles in a data set of ten observations are the third and eighth observations, it's possible to find several data sets with an interquartile range of 1. Figure 3.4 illustrates two possible teams that meet this criteria.

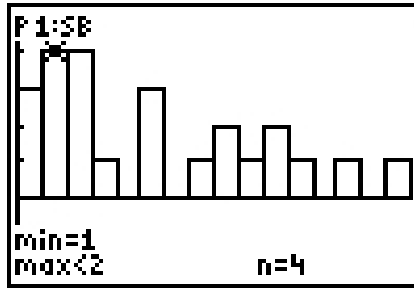
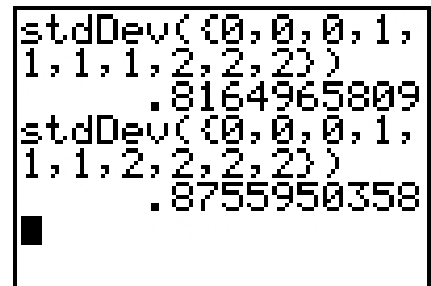


FIGURE 3.5. Visualizing Part 5 with a Histogram

For the minimum standard deviation, we recognize that the best team in this sense has the “tightest” set of stolen base values. We can visualize the ordering of players by stolen bases by plotting the data. A dotplot is requested in Part 5, but the TI-83 does not have a dotplot option. Here we substitute a histogram of column width one which produces roughly the same image. Figure 3.5 shows the histogram of stolen base data after the WINDOW has been adjusted to ignore the outlier of 41 stolen bases. We can now see that we want to choose a team from the twelve players with 0, 1, 2 or 3 stolen bases.

Number of 0's	Number of 1's	Number of 2's	Number of 3's	Mean	Standard Deviation
3	4	3	0	1.0	0.8165
3	3	4	0	1.1	0.8756
<b>2</b>	<b>4</b>	<b>4</b>	<b>0</b>	<b>1.2</b>	<b>0.7888</b>
1	4	4	1	1.5	0.8498
2	4	3	1	1.3	0.9487
2	3	4	1	1.4	0.9661
3	4	2	1	1.1	0.9944
3	3	3	1	1.2	1.0328
3	2	4	1	1.3	1.0593

(a) Table of Possible Team Combinations



(b) Comparing Standard Deviations

FIGURE 3.6. Answering Part 5

To determine which collection gives the smallest standard deviation, we use the `stdDev` function with different stolen base combinations. Figure 3.6 illustrates the possible combinations for a ten-player team. The team that produces the smallest standard deviation includes all players with 1 and 2 stolen bases and 2 players with 0 stolen bases.

**3.2. Hypothesis Test with the Smallest  $p$ -Value.** Although it may not appear to be a question that can be solved with a simple sorting of data, an understanding of hypothesis tests enables Part 9 to be answered in just that manner. Part 9 challenges us to find a ten-player team that results in the smallest  $p$ -value for

the hypothesis test

$$H_0 : \pi_{\text{highAVG}} = .5 \text{ vs. } H_A : \pi_{\text{highAVG}} > .5,$$

where  $\pi_{\text{highAVG}}$  is the proportion of batting averages in MLB that are greater than .270.

The test statistic is:

$$(3.1) \quad z = \frac{p_{\text{highAVG}} - .5}{\sqrt{\frac{(.5)(.5)}{10}}}$$

where  $p_{\text{highAVG}}$  is the proportion of batting averages greater than .270 for our ten-player team. For this upper-tailed hypothesis test, the  $p$ -value is the area under the standard normal curve that is illustrated in Figure 3.7.

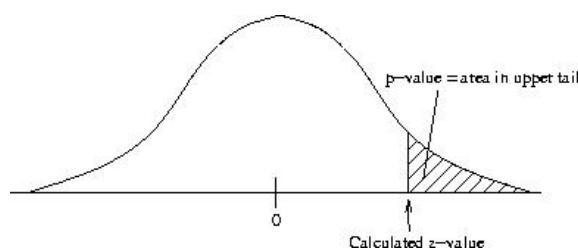


FIGURE 3.7. Illustrating the  $p$ -value

To minimize the  $p$ -value, we need to find the largest possible positive  $z$ -value. Since all components of this test statistic are constant except the value of  $p_{\text{highAVG}}$ , maximizing the  $z$ -value means choosing the team with the largest  $p_{\text{highAVG}}$  value. Thus we simply need to sort the data by **AVG** and select the ten players with the highest batting average. Note, this data set has 15 players with a batting average greater than .270. The smallest  $p$ -value is achieved using any ten of these players. Thus this question has 3,003 ( ${}_{15}C_{10}$ ) correct answers.

**3.3. Strong Linear Relationships Between Runs and Home Runs.** Determining a team with a strong linear relationship between runs and home runs (Part 6) relies on a scatter plot (see Figure 3.8) and an understanding of correlation. Note, the project does not require identification of the team with *the* strongest linear relationship, as this is not easily identified visually, and comparing the correlation coefficients for each of the 3,268,760 ( ${}_{25}C_{10}$ ) possible teams is computationally intensive. Instead, we look at the scatter plot and try to identify ten points that are highly collinear. Note, this does not mean finding the ten points that lie closest to the least squares regression line for the entire data set, yet some students take this approach.

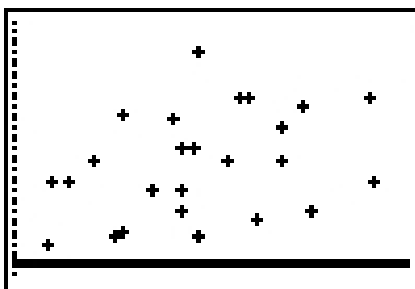


FIGURE 3.8. Looking for Linear Relationships in a Scatter Plot of Runs versus Home Runs

ORDER	TEAMR	TEAMH
1	68	11
2	66	3
3	65	15
4	63	2
5	61	11
6	51	4

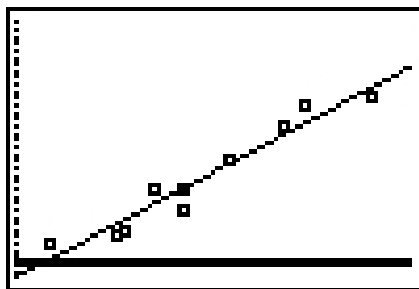
TEAMR = {11, 3, 15, 2...

(a) Selected Team Data

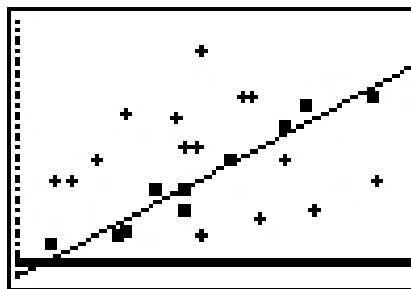
```

LinReg
y=ax+b
a=.3157915276
b=-10.30541084
r=.9319420283
r=.9653714458
  
```

(b) Linear Regression Results



(c) Least Squares Line with Team



(d) Least Squares Line with All Data

FIGURE 3.9. Quantifying and Illustrating the Linear Relationship Between Runs and Hits

Once we have selected a team of ten players that we feel demonstrates a strong linear relationship between their runs and home runs (the TRACE feature on the TI-83 will help with this process), we can calculate the associated correlation coefficient. It is easiest to create new lists containing the run and home run data for the ten-player team we selected, then perform the linear regression on this smaller set of data. Figure 3.9 illustrates this process: entering the team data 3.9(a), determining the correlation coefficient for this data 3.9(b), and finally plotting the least squares line with the team data 3.9(c) and the original set of data 3.9(d).

#### 4. EXTENDING THIS PROJECT

Which of the  ${}_{25}C_{10}$  ten-man teams is the best overall? Most students will choose a team that they think is best, but will not prove that their team is best. A student who enjoys programming could write a computer program that cycles through all possible 10-man teams and determines which one is best. Is there a less exhaustive approach? Perhaps this could be your next “team” project.

## APPENDIX A. PLAYER DATA

Player Number	First Name	Last Name	AB	H	R	SB	SO	HR	RBI
1	Angel	Berroa	608	164	68	7	108	11	55
2	Aramis	Ramirez	463	140	72	0	60	31	92
3	Brad	Ausmus	387	100	35	5	48	3	47
4	Brian	Giles	545	164	92	13	64	15	83
5	Darin	Erstad	609	166	86	10	109	7	66
6	Edgar	Renteria	623	172	100	9	100	8	70
7	Eric	Hinske	477	125	79	8	121	15	68
8	Felipe	Lopez	580	169	97	15	111	23	85
9	Freddy	Sanchez	453	132	54	2	36	5	35
10	Garret	Anderson	575	163	68	1	84	17	96
11	Gregg	Zaun	434	109	61	2	70	11	61
12	Jeromy	Burnitz	605	156	84	5	109	24	87
13	Jerry	Hairston	380	99	51	8	46	4	30
14	Jhonny	Peralta	504	147	82	0	128	24	78
15	Jimmy	Rollins	677	196	115	41	71	12	54
16	Joe	Crede	432	109	54	1	66	22	62
17	John	Buck	401	97	40	2	94	12	47
18	Juan	Rivera	350	95	46	1	44	15	59
19	Luis	Castillo	439	132	72	10	32	4	30
20	Michael	Young	668	221	114	5	91	24	91
21	Nick	Swisher	462	109	66	0	110	21	74
22	Ramon	Hernandez	369	107	36	1	40	12	58
23	Ronnie	Belliard	536	152	71	2	72	17	78
24	Russ	Adams	481	123	68	11	57	8	63
25	Todd	Helton	509	163	92	3	80	20	79

## APPENDIX B. SOLUTIONS WITH THE TI-83

## Abbreviated Solution Procedures for the Fantasy Baseball Project Using the TI-83

Part	Team Objective	Solution Approach on TI-83
1	Highest Mean RBI	<p>(1) Sort the data by RBI.</p> <p>(2) Pick players with the ten highest RBI values.</p>
2	Smallest Mean SO	<p>(1) Sort the data by SO.</p> <p>(2) Pick players with the ten lowest SO values.</p>
3	Highest Median HR	<p>(1) Sort the data by HR.</p> <p>(2) Pick players with the ten highest HR values.<sup>3</sup></p> <hr/> <p><sup>3</sup>Note, this is just one way to end up with the highest median HR value. You should choose the players with the six highest HR values and <i>any</i> four of the remaining players.</p>
4	Highest Median AVG	<p>(1) Sort the data by AVG.</p> <p>(2) Pick players with the ten highest AVG values.<sup>4</sup></p> <hr/> <p><sup>4</sup>See footnote <sup>3</sup> above.</p>
5	Minimum Variation SB	See Section 3.1.

Part	Team Objective	Solution Approach on TI-83
6	Largest Magnitude Correlation Coefficient, $ r $ , for R and HR	See Section 3.3
7	Smallest $P(SB < 10)$	<p>(1) Sort the data by SB.</p> <p>(2) Pick players with the ten highest SB statistics.<sup>5</sup></p> <hr style="width: 20%; margin-left: 0;"/> <p><sup>5</sup>Note, this is just one way to end up with the smallest <math>P(SB &lt; 10)</math> value. You should choose the six players with SB values greater than or equal to 10, and <i>any</i> four of the remaining players.</p>
8	Narrowest 95%-Confidence Interval for $\mu_{AVG}$	<p>Follow the directions for picking a team with the smallest standard deviation in stolen bases, but use batting average as the ordering variable. The 95%-confidence interval for <math>\mu_{AVG}</math> is given by</p> $\bar{x}_{AVG} - 2.06 \frac{s_{AVG}}{\sqrt{10}} \leq \mu_{AVG} \leq \bar{x}_{AVG} + 2.06 \frac{s_{AVG}}{\sqrt{10}},$ <p>thus the narrowest confidence interval results from having the smallest sample standard deviation, <math>s_{AVG}</math>.</p>
9	Smallest $p$ -value for the Hypothesis Test $H_0 : \pi_{\text{highAVG}} = .5$ vs. $H_A : \pi_{\text{highAVG}} > .5$	See Section 3.2

## REFERENCES

- [1] Fantasy Baseball. Wikipedia, The Free Encyclopedia: [http://en.wikipedia.org/wiki/Fantasy\\_baseball](http://en.wikipedia.org/wiki/Fantasy_baseball).

- [2] The College Board Advanced Placement Program: Statistics Course Description, May 2005 - May 2006. The College Board Web Site, [http://apcentral.collegeboard.com/repository/statistics\\_cd.0506\\_4328.pdf](http://apcentral.collegeboard.com/repository/statistics_cd.0506_4328.pdf), 2004.
- [3] Roxy Peck, Chris Olsen, and Jay Devore. *Introduction to Statistics and Data Analysis, 2<sup>nd</sup> Edition*. Thomson Brooks/Cole, 2005.