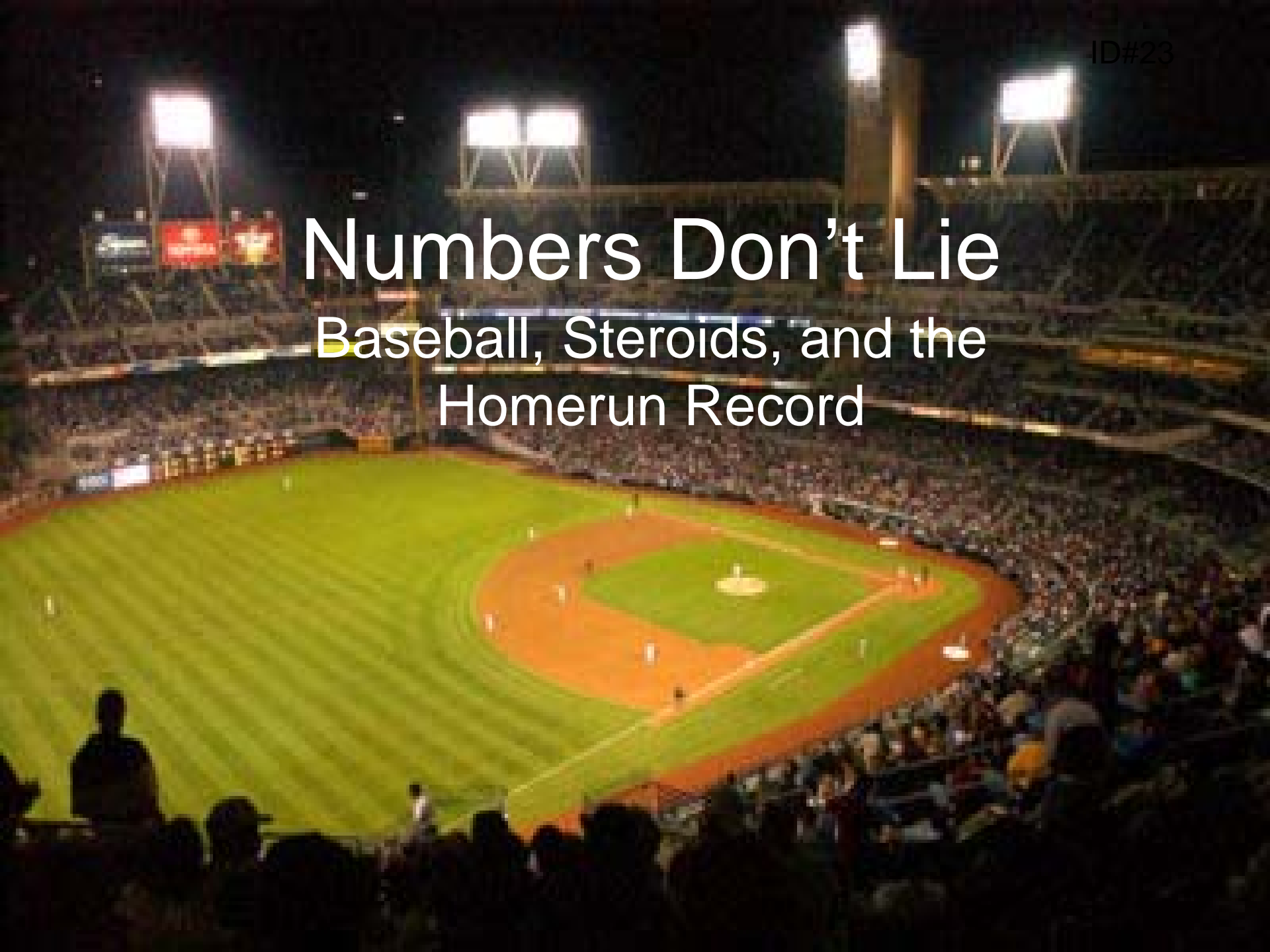


# Numbers Don't Lie

## Baseball, Steroids, and the Homerun Record





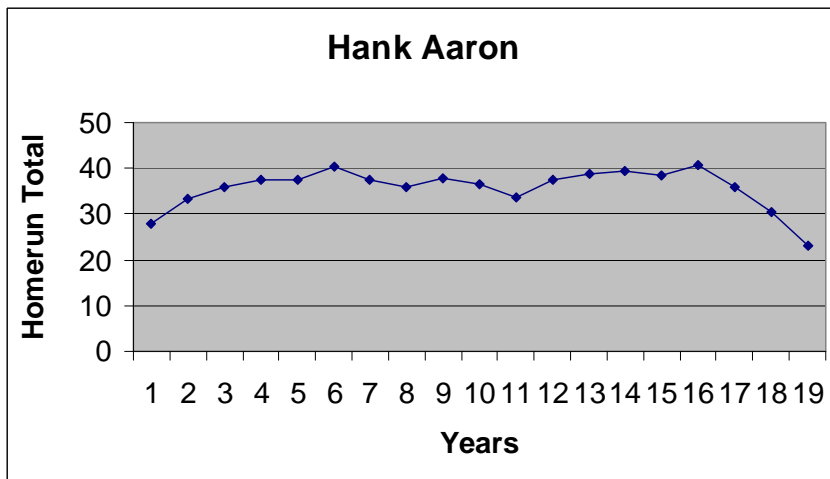
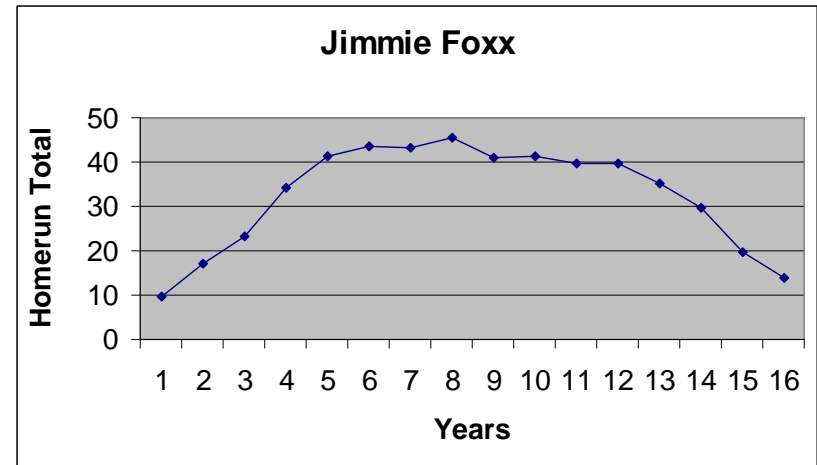
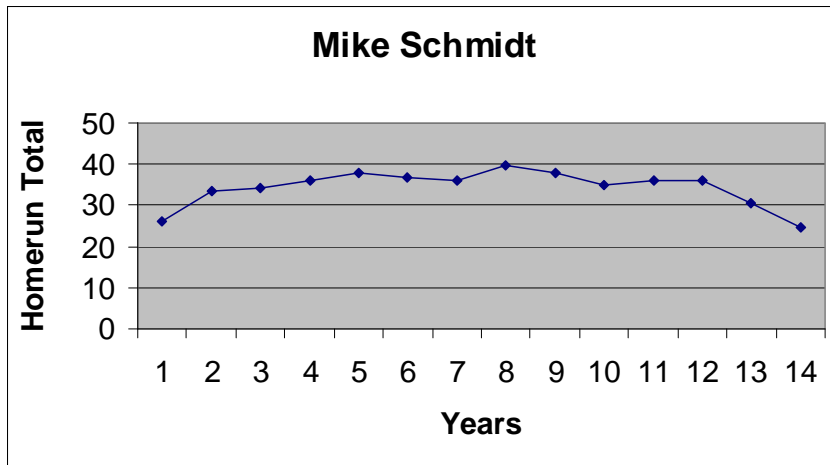
# Introduction

- Statistics are as much a part of the game of baseball as the bat and the ball.
- Today's homerun statistics are under suspicion as a result of the introduction of steroids.
- No debate surrounding the homerun records can go without the mention of steroids.
- The question is, how much of an impact have steroids truly had?
- With all of the recent allegations, is there a way to statistically determine who may or may not have utilized steroids?

# Background Information

- There are three groups to this study
  - Non-users
    - These are players who with almost 100% certainty did not use steroids. They either played before steroids were introduced, or were cleared by testing.
  - Assumed Users
    - These are players who do not have specific evidence proving whether they did or did not use steroids. However, these players have been accused by their peers as steroid users, but have not been caught.
  - Users
    - These are players who have either admitted to, or have been caught through testing using steroids.
- Unless otherwise specified, throughout the course of this study, the group “Users” will also include “Assumed Users.”
- Steroid years refer to the seasons which are known or assumed that the player used steroids.
- Throughout the study, the term year (shown by 1,2,3 etc. on graphs and charts) refers to the first, second, third, etc. year of a player’s career.

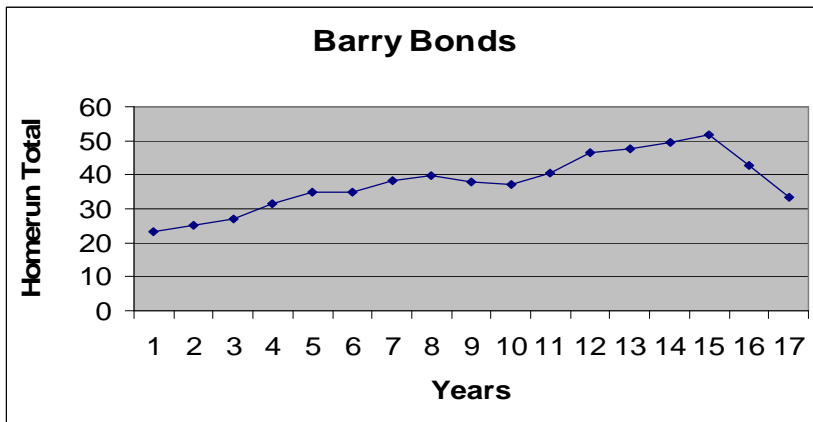
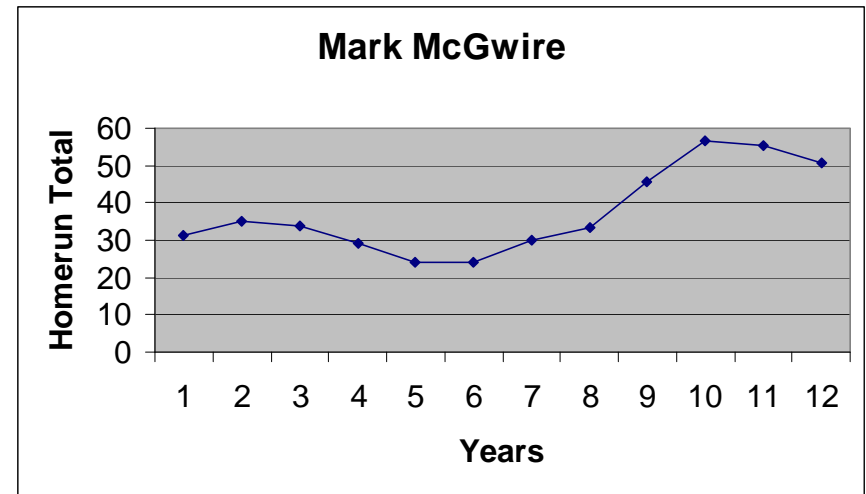
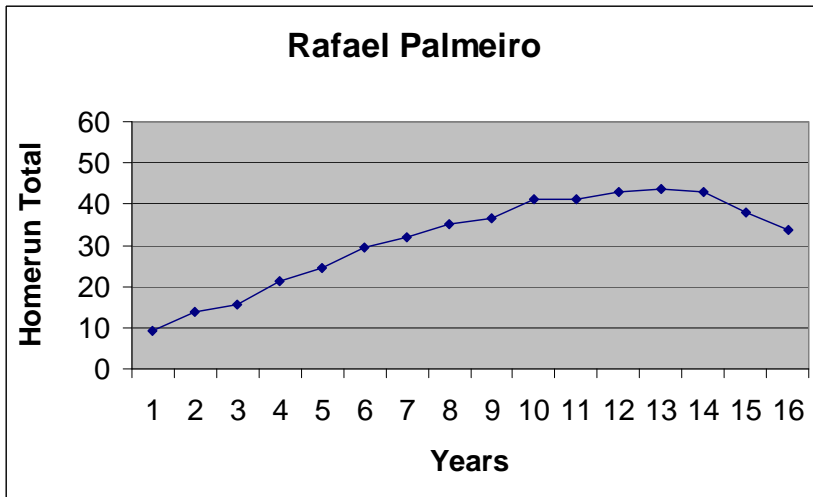
# Non-Users



- Graphs show 5 year moving averages
- Players who do not use steroids tend to show a more symmetrical rise and fall to their career homeruns
- They tend to peak in the middle of their career

\*This is a selection of three players, although the vast majority of non-users show this type of curve.

# Steroid Users

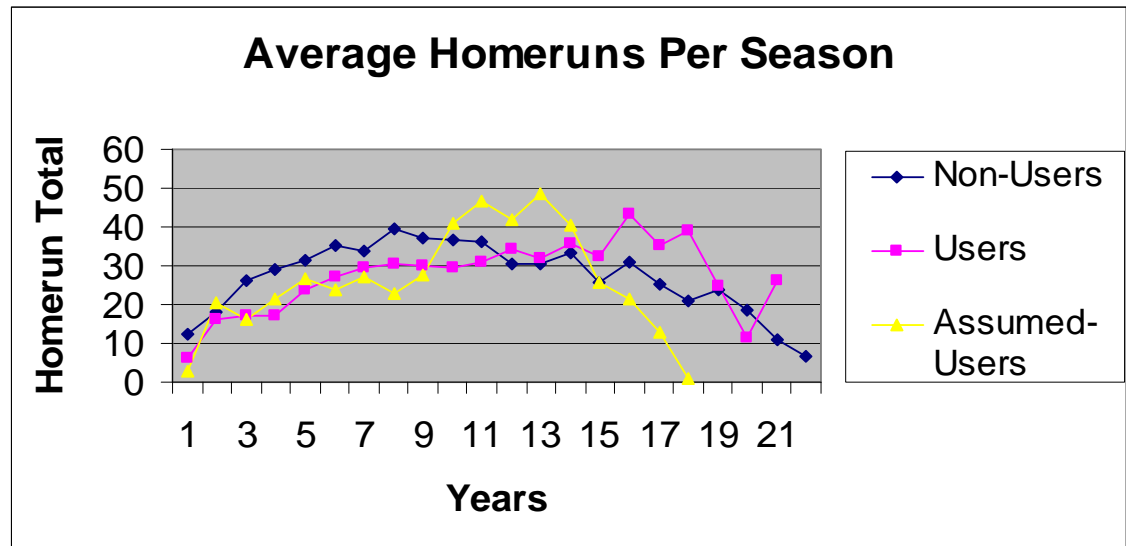


- Graphs show 5 year moving averages
- Tend to follow a “hockey stick” shaped distribution
- These players tend to have a constant rise until late in their careers when they reach their peaks

\*This is a selection of three players, although the vast majority of users show this type of curve.

# Homerun Averages

Average Homerun Totals	
Average Before Steroids	19.29
Average During Steroids	31.42
Average After Steroids	22.38
Average Never Using Steroids	28.13



The graph shows the non-users, users, and assumed-users averages for each year of their careers

Notice that non-users start out their careers hitting more homeruns on average than eventual users. These low early career totals may be the reason why certain players turn to steroids.

Steroid Users and Assumed Users					Highest Homerun Total			
		Age First Used		Age Stopped Using		With Steroids		Without Steroids
Jose Canseco		20		36		46		N/A
Rafael Palmeiro		27		40		47		26
Mark McGwire		24		37		70		49
Sammy Sosa		20		36		66		N/A
Barry Bonds		34		38		73		46
Ken Caminiti		33		38		40		26
Ivan Rodriguez		20		31		35		19
Gary Sheffield		30		34		43		42
Jason Giambi		26		32		43		37
Andres Galarraga		32		43		47		29
Mean		26.6		36.5		51		34.25
Median		26.5		36.5		46.5		33

- Average number of homeruns for steroid users ages 26-36 = 31.95
- Average number of homeruns for non-users ages 26-36 = 32.29

# Average of 5 Best Years

Sammy Sosa	58.4
Mark McGwire	57.4
Babe Ruth	55.2
Ken Griffey Jr.	50.8

Barry Bonds	49.4
Jimmie Foxx	48.2
Willie Mays	48.0
Harmon Killebrew	47.4
Mickey Mantle	45.0
Hank Aaron	44.8
Ernie Banks	44.0
Rafael Palmeiro	43.8
Eddie Matthews	42.6
Mike Schmidt	41.8
Willie McCovey	40.6
Jose Canseco	40.6

Frank Robinson	39.8
Jason Giambi	39.2
Reggie Jackson	39.0
Andres Galarraga	38.8
Ted Williams	38.4
Mel Ott	36.8
Gary Sheffield	35.4
Eddie Murray	31.6
Ivan Rodriguez	25.6
Ken Caminiti	25.0

These are the average homerun totals for each player's five best seasons. In red are the steroid users, and their average is found by using only their steroid numbers.

Steroid users average of 5 best years = 41.36

Non-users average of 5 best years = 43.38

# Expected Homeruns in a Single Season

- Non-users expected homeruns = 32.92
- Users expected homeruns = 35.53
  - These expected values were found by:
    - Breaking down the homerun totals into increments of 10 (i.e. a single season total of 7 falls into category of 0-9, and 38 falls into 30-39 etc.)
    - The frequency of each category was divided by the total number of seasons. This value became the weights.
    - These weights were then multiplied by their corresponding homerun totals (5, 15, 25, etc).
    - These values were then added together to create the expected value.
    - An example of this process is shown on the next slide.

# Expected Value Example

- Given three players single season totals of: player 1 (12, 23,21), player 2 (5,8) and player 3 (15,14).
- These totals would fall into the following categories:
  - 0-9 (5,8) for a frequency of 2
  - 10-19 (12,14,15) for a frequency of 3
  - 20-29 (21,23) for a frequency of 2
- The total number of seasons is 7 so our weights are:
  - 0-9 =  $2/7$  with a corresponding total of 5
  - 10-19 =  $3/7$  with a corresponding total of 15
  - 20-29 =  $2/7$  with a corresponding total of 2
    - The corresponding total is the median value of each category.
- The expected value =  $5*(2/7) + 15*(3/7) + 25*(2/7) = 15$

# Two-Sample Test of Significance

- By comparing the weighted average of homeruns for users and non-users, it became apparent that no conclusions could be drawn from those two numbers alone.
  - Weighted average for non-users = 28.10
  - Weighted average for users = 27.23
- The previous data shows a striking difference between the first and second half careers of users compared to non-users.
  - The first half weighted average for users = 21.49
  - The first half weighted average for non-users = 30.19
  - The second half weighted average for users = 32.86
  - The second half weighted average for non-users = 26.31
- This was enough evidence to warrant a significance test to determine whether or not the users' population mean was lower than the non-users' population mean in the first half of their careers. This will be called the first-half test.
- Likewise, another significance test would determine whether or not the users' population mean was higher than the non-users' population mean in the second half of their careers. This will be called the second-half test.

# Two-Sample Test of Significance

- In order to find the significance of the data, the mean, standard deviation, and sample size was found for each half of the career of both users and non-users.

	First-half Test	Second-half Test
Sample Mean users	21.49	32.86
Sample Standard Deviation users	11.44	13.14
Sample Size users	88	88
Sample Mean non-users	30.19	26.31
Sample Standard Deviation non-users	6.11	8.92
Sample Size non-users	165	165

A two-sample t-test was executed and the following p-values were found:

- The first-half test =  $5.4 \times 10^{(-10)}$
- The second-half test =  $2.55 \times 10^{(-5)}$

# Conclusion of Significance Test

- Since the p-value of the first-half test is less than .01, there is very strong statistical evidence that the users' population mean is significantly lower than the non-users' population mean in the first-half of their careers.
- Since the p-value of the second-half test is less than .01, there is very strong statistical evidence that the users' population mean is significantly higher than the non-users' population mean in the second-half of their careers.
- Based on the significance tests, while the non-users' homerun totals are declining in the second half of their careers, the steroid users' totals are increasing.

# Breaking the Record

- The current single season homerun record is 73 set by Barry Bonds in 2003. In 2003 Barry Bonds admitted to using steroids.
- The highest homerun total by any player not on steroids or not assumed to use steroids is 61 by Roger Maris in 1961.
- What is the probability that the current record of 73 can be broken by a player not on steroids and by a player on steroids?
  - By making a histogram of each non-users' and users' homerun data, and fitting a cubic regression an equation can be found that will roughly estimate the probability of hitting 74 homeruns.
    - The regression equation for non-users:  
 $f(x) = 0.9562 + 0.6275x - 0.017x^2 + 0.0001x^3$
    - The regression equation for users:  
 $f(x) = 1.293 + 0.2896x - 0.0089x^2 + 0.000065x^3$
- The probability that a non-user will break the record is 0 using the data given.
- The probability that a user will break the record is 0.001871

# Conclusion

- Over the course of a full career, steroids do not seem to enhance homerun hitting. However, for a single season a player may hit more homeruns while using steroids.
- Non-users seem to have their peak around the middle of their career, while users tend to peak later in the second half of their career.
- Based on our data we were able to determine with relative confidence whether or not the four assumed users did in fact use steroids.
  - Andres Galarraga, Sammy Sosa, and Mark McGwire fit the characteristics of a typical steroid user.
  - Ivan Rodriguez's career does not fit the model of a user, but in fact fits the model of a non-user almost perfectly.
  - This model will only work for players who are the top homerun hitters in the game, since our sample is from the greatest of all time. This study must also be done at the end of a player's career to get the full extent of the model.

# Sources

- [www.baseball-reference.com](http://www.baseball-reference.com)
- Canseco, Jose. Juiced: Wild Times, Rampant Roids, Smash Hits, and How Baseball Got Big. 2005 ReganBooks. New York, NY.
- <http://thesteroidera.blogspot.com>
- [www.baseball-almanac.com](http://www.baseball-almanac.com)
- [www.baseballhalloffame.org](http://www.baseballhalloffame.org)