INTRO TO SABERMETRICS AND STATISTICS

WHAT IS SABERMETRICS?

- "The search for objective knowledge about baseball." – Bill James
- The term was coined, in part, to honor the Society for American Baseball Research (SABR)
- Sabermetric researchers often use statistical analysis to question traditional measures of baseball evaluation such as batting average and pitcher wins
- "The main contribution of Bill James and Pete Palmer … is their exposure of the deficiency of looking at merely the traditional statistics." – "Understanding Sabermetrics" by Gabriel B. Costa, Michael R. Huber and John T. Saccoman

THE MEAN

- What is the mean?
- The mean, or average is the sum of all elements divided by the total number of elements in the set
- Practicing the Mean: Let's look at yearly home run totals for Albert Pujols
- 37; 34; 43; 46; 41; 49; 32; 37; 47; 42; 37; 30; 17; 23
- What is the the mean?
- 515 home runs / 14 seasons =
- 36.8

MEDIAN AND MODE

- What is the Median?
- The median is the exact middle of the data set
- Pujols' home runs (in order):
  - 17; 23; 30; 32; 34; 37; 37; 37; 41; 42; 43; 46; 47; 49
- What is the mode?
- The mode is the number that appears most frequently.
- What is Pujols' mode and median?

EXPECTATIONS

- Pujols currently has 515 home runs. Looking at his average output, how many more seasons does he need to play to hit 600 career home runs?

MEASURES OF DISPERSION

- In statistics, measures of dispersion show how tightly spread out the data is in relation to a measure of central tendency.
- The main measures of dispersion are: Range, Variance and Standard Deviation
- Range: Pujols' season totals vary from 17 to 49, giving him a range of (maximum – minimum = range) 32
- Broken into quartiles, the first quartile would be values less than 32; the second quartile ends at the median (37) and the third quartile ends at 43.
- The interquartile range (IQR) is 43 – 32= 11, meaning that 50 percent of the data is separated by 11.

1st Quar.	2nd Quar.	3rd Quar.	4th Quar.
17; 23; 30; 32;	34; 37; 37;	37; 41; 42; 43;	46; 47; 49

VARIANCE AND STANDARD DEVIATION

- s^2 = Variance
- Σ = Summation, which means the sum of every term in the equation after the summation sign.
- x_i = Sample observation. This represents every term in the set.
- $\overline{x}$ = The mean. This represents the average of all the numbers in the set.
- n = The sample size. You can think of this as the number of terms in the set.
- Put simply: The variance is the sum of the squared difference between each number and the mean, divided by the total items in the set.

$$s^2 = \frac{\sum (X_i - \overline{X})^2}{n}$$

PUJOLS' HOME RUN VARIANCE

- Pujols' Home Runs:
  - 37; 34; 43; 46; 41; 49; 32; 37; 47; 42; 37; 30; 17; 23
- The Mean: 36.8 ($\overline{x}$)
- Take each difference, square it and average the result
- $(.2)^2+(-2.8)^2+(6.2)^2+(9.2)^2+(4.2)^2+(-4.8)^2+(.2)^2+(10.2)^2+(5.2)^2+(.2)^2+(-6.8)^2+(-19.8)^2+(-13.8)^2$
- .04+7.84+38.44+84.64+17.64+23.04+.04+104.04+27.04+.04+46.24+392.04+190.44 = 931.52
- Variance = 931.52/14= 66.54
- To find the standard deviation, take the square root of the variance
- Standard Deviation = 8.16

CHEBYSHEV'S RULE

- According to Chebyshev's Rule – for a data set that doesn't follow a bell curve – at least 95 percent of the data will fall within 2 Standard Deviations of the Mean, and at least 99.4 percent will fall within 3.
- Looking at Pujols' Home Run Totals again:
  - Mean: 36.8
  - Standard Deviation: 8.16
  - So, 95% of his home run totals should be: 20.48 to 53.12. In fact, just one year falls outside of that range: 2013's 17.