Baseball with Popcorn, Statistics and SAS - What A Mix

Paper SA05

David Franklin
TheProgrammersCabin.com, Independent SAS Consultant

While you are waiting, some trivia ….
"Q" is the only letter in the alphabet that does not appear in the name of any state of the United States.

Introduction

- Motivation: a local little league wanted to know who was the best batter and best pitcher for the season
- Paper looks at (from 30,000 feet) how baseball data are collected (in paper form), how some statistics are calculated, and finally presents a program that was used to find the best batter and best pitcher for the season

How is the Raw Data Collected

Team:	vs	at	www.baseballscorecard.com
Date:	Start Time:	End Time:	Time of Game:
Weather:			
Umpires:		Scorer:	

Batting

#	Player	Pos	1	2	3	4	5	6	7	8	9	10	AB	R	H	RBI	E
	sub																
	sub																
	sub																
	sub																
	sub																
	sub																
	sub																
	sub																
	sub																
	sub																

		1	2	3	4	5	6	7	8	9	10	TOTALS	
S	Runs												
U	Hits											AB:	RBI:
M	Errors											R:	E:
S	Left on Base											H:	LOB:

Pitching

Pitcher	W-L	IP	H	R	ER	BB	SO	HB	BK	WP	TBF	Catcher	PB

http://www.baseballscorecard.com

What is Collected on the Form I

- There are many versions of the form – the one shown on the previous slide is just one of many.
- Batting:
  - AB (At bat): Batting appearances, not including bases on balls, hit by pitch, sacrifices, interference, or obstruction.
  - R (Runs scored): number of times a player crosses home plate
  - H (Hits): times reached base because of a batted, fair ball without error by the defense
  - RBI (Run batted in): number of runners who scored due to a batters' action, except when batter grounded into double play or reached on an error
  - E (Error): an act where a fielder misplaying a ball in a manner that allows a batter or base runner to reach one or more additional bases

What is Collected on the Form II

- Pitching:
  - IP (innings pitched)
  - H (Hits): total hits allowed
  - R (Runs): total runs allowed
  - ER (earned runs): number of runs that did not occur as a result of errors or passed balls
  - BB (walks): times pitching four balls, allowing the batter-runner to advance to first base
  - SO (strikeouts)
  - HB (batters hit): times hit a batter with pitch, allowing runner to advance to first base
  - BK (balks): number of times pitcher commits an illegal pitching action or other illegal action while in contact with the pitching rubber, thus allowing base runners to advance
  - WP (wild pitches): charged when a pitch is too high, low, or wide of home plate for the catcher to field, thereby allowing one or more runners to advance or score
  - TBF (batters faced): opponent's total plate appearances

What is Collected on the Form III

- Note that other forms of the score card may collect other forms of information including:
  - number of Intentional Walks
    - number of Inherited Runs Allowed
    - Outfield Assists
    - Passed Balls
    - Pickoffs
    - Putouts

This is commonly beyond the scope of minor league statistics.

Statistics

- Over 150 statistics can be calculated! Common ones are:
  - Batting Average: (total hits) / (times at bat)
  - Earned Run Average: 9 * (earned runs) / (innings pitched)
  - Base-on-balls Percentage: (total walks) / (plate appearances)
  - Home Run Ratio: (at-bats) / (home runs)
  - On-base Percentage: (hits + walks + hits by pitch) / (at-bats + walks + hits by pitch + sacrifice flies)
  - Runs Created: [(hits + walks - caught stealing) * (total bases + (stolen bases * 0.55)] /(at-bats + walks)
  - Slugging Average: (total bases) / (at-bats)
  - Won-Lost Percentage: (wins) / (wins + losses)
  - Fielding Average: (total putouts + assists) / (putouts + assists + errors)
  - Fielder's Range Factor: (putouts + assists) / (games)
  - Opponents' Batting Average: (hits allowed) / (at bats allowed)
  - Strikeout Ratio: (at-bats) / (strikeouts)
  - Stolen Base Percentage: (stolen bases) / (total attempts)
  - Winning Percentage: (games won) / (total games played)
- Most can be calculated off the form, others need a modified version

Our Real World Example I

- Two datasets, one for Batting and the second for pitching
- Data comes from Excel spreadsheets that each coach uses to enter the data into the SAS program (reference Heaton, E. Reading Excel® Workbooks. SAS Global Forum, 119-2007, 2007)
- Batting Dataset (SAS variable names in [])
    - Team [TEAM]
    - Player Number [PLAYER]
    - Date [DATE]
    - Start Time (if a team had multiple games on the same day, useful if trying to get individual information) [TIME]
    - Number of Times At bat [NUMBAT_B]
    - Runs Scored [RUNS_B]
    - Hits [HITS_B]
    - RBI [RBI_B]
    - Errors [ERRORS_B]

Our Real World Example II

- The Pitching Dataset (SAS variable names are in [])
  - Team [TEAM]
  - Player Number [PLAYER]
  - Date [DATE]
  - Start Time (if a team had multiple games on the same day, useful if trying to get individual information) [TIME]
  - Innings Pitched [INPTCH_P]
  - Hits [HITS_P]
  - Runs [RUNS_P]
  - Earned Runs [EARNRN_P]
  - Walks [WALKS_P]
  - Strikeouts [STRICK_P]
  - Batters Hit [BATHIT_P]
  - Balks [BALKS_P]
  - Wild Pitches [WILD_P]
  - Batters Faced [BATFCE_P]

Our Real World Example III

Dataset: BATTING

TEAM	PLAYER	DATE	TIME	NUMBAT_B	RUNS_B	HITS_B	RBI_B	ERRORS_B
Giants	4	4-May-11	17:00	4	2	4	0	1
Giants	6	4-May-11	17:00	5	1	2	0	1
A's	8	4-May-11	17:00	4	0	1	0	0
A's	15	4-May-11	17:00	5	4	5	1	0
Diamondbacks	3	5-May-11	17:00	4	0	1	0	1
Diamondbacks	7	5-May-11	17:00	4	1	2	0	0
Cubs	4	5-May-11	17:00	4	0	2	0	0
Cubs	5	5-May-11	17:00	4	0	1	0	0

Dataset: PITCHING

TEAM	PLAYER	DATE	TIME	INPTCH_P	HITS_P	RUNS_P	EARNRN_P	WALKS_P
Giants	7	4-May-11	17:00	3	6	4	4	2
Cubs	13	5-May-11	17:00	2	4	6	1	0

Dataset: PITCHING

TEAM	PLAYER	DATE	TIME	STRICK_P	BATHIT_P	BALKS_P	WILD_P	BATFCE_P
Giants	7	4-May-11	17:00	4	0	-	-	20
Cubs	13	5-May-11	17:00	5	0	-	-	17

Our Real World Example IV

The Calculations

- Batting Average: (total hits) / (times at bat)
- Used the SUMMARY procedure for summation across season, then the RANK procedure to compute the ranking of the player – due to its options an analysis by team and overall league could be computed
- With the Batting Average, the higher the value the better!

- Earned Run Average: 9 * (earned runs) / (innings pitched)
- Again, used the SUMMARY procedure for summation across season, then the RANK procedure to compute the ranking of the player – due to its options an analysis by team and overall league could be computed
- With the Earned Run Average, the lower the value the better!

Our Real World Example V

- Code is in the paper – too long to produce here
- Output from sample:

Batting Average for League
Based on Data Received by 07May2011

Rank	Team	Player	Batting Average	Number of Hits	Number of Times at Bat
1	A's	15	1.000	5	5
1	Giants	4	1.000	4	4
3	Cubs	4	0.500	2	4
3	Diamondbacks	7	0.500	2	4
5	Giants	6	0.400	2	5
6	A's	8	0.250	1	4
6	Cubs	5	0.250	1	4
6	Diamondbacks	3	0.250	1	4

Pitching Average for League
Based on Data Received by 07May2011

Rank	Team	Player	Earned Run Average	Earned Runs	Innings Pitched
1	Cubs	13	4.500	1	2
2	Giants	7	12.00	4	3

About the RANK Procedure

- The RANK procedure was used to put a numerical “rank” against the data and it dealt easily with tied values
- The TIES= option controls the treatment of tied values. Possible values for this option are:
  - LOW: assigns the smallest of the corresponding ranks to these observations (if values are 0.1, 0.2, 0.2 and 0.3 then ranks will be 1, 2, 2 and 4)
  - HIGH: assigns the largest of the corresponding ranks (if values are 0.1, 0.2, 0.2 and 0.3 then ranks will be 1, 3, 3 and 4)
  - MEAN: assigns the mean of the corresponding rank (if values are 0.1, 0.2, 0.2 and 0.3 then ranks will be 1, 2.5, 2.5 and 4).
- In each of these cases the values are treated as distinguishable values.
- In recent versions of SAS, the option TIES=DENSE has been available that computes scores and ranks by treating tied values as a single-order statistic, i.e. tied values are treated as indistinguishable with each value within a tied group is assigned the same ordinal.

Conclusion

- The question asked in the local minor league was who is the best batter and who is the best pitcher, was answered.
- Using SAS, the league was able to compute and easily report the result.
- Because of the number of data points collected it is also possible to do more statistical computations and there is an idea for the fall season to have this report available to the league officials after half-way though the season on a week by week basis.
- Beyond the original scope of the specification, best pitcher and best batter will be calculated by team as well as across the league

Questions and Contact Information

Questions?

Contact Information
David Franklin
TheProgrammersCabin.com
16 Roberts Road, Litchfield, NH 03052
Phone: 603-275-6809
Email: dfranklin@TheProgrammersCabin.com
Web: http://www.TheProgrammersCabin.com
LinkedIn: http://www.linkedin.com/in/davidfranklinnh

Acknowledgement