Basic Descriptive Statistics Analysis on Baseball

Ruixie Fang

Data Description

This dataset contains information for 337 Major League Baseball (MLB) players who are not pitchers and played at least one game during both the 1991 and 1992 seasons, such as the players' 1992 salaries, batting average, number of home runs and other various measures of the players' 1991 performance.

Number Of Records

337 observations

18 variables (full_name-character; avg,OBP-continuous; others-categorical)

Source

Sacramento Bee October 15, 1991

New York Times November 13, 1991/ November 19, 1992/ February 23, 1992.

CNN/Sports Illustrated at

http://www.cnnsi.com/baseball/mlb/historical_profiles/

The Society for American Baseball Research (SABR) at

ftp://skypoint.com/pub/members/a/ashbury/sabr/SALARIES/1992_salaries_baseball

Basic Descriptive Statistics

The purpose of this analysis is to determine whether a baseball player's salary is based on his performance. First, summary information is processed by using PROC CONTENTS statement. The output describes the structure of the data set, including number of records and description of variables (Table 1).

Table 1. Summary Contents About Dataset

Data Set Name	WORK.BASEBALL	**Observations**	337
Member Type	DATA	**Variables**	18

Table 2. Description Of Variable

Variables in Creation Order				
#	**Variable**	**Type**	**Len**	**Label**
1	salary	Num	8	Salary (in thousands of dollars)
2	avg	Num	8	Batting average
3	OBP	Num	8	On-base percentage (OBP)
4	Runs	Num	8	Number of runs
5	Hits	Num	8	Number of hits
6	Doubles	Num	8	Number of doubles
7	Triples	Num	8	Number of triples
8	HRs	Num	8	Number of home runs
9	RBI	Num	8	Number of runs batted in (RBI)
10	Walks	Num	8	Number of walks
11	SOs	Num	8	Number of strike-outs
12	SBs	Num	8	Number of stolen bases
13	Errors	Num	8	Number of errors
14	FA_Eligible	Num	8	Indicator of "free agency eligibility"
15	FA_9192	Num	8	Indicator of "free agent in 1991/2"
16	Arb_Eligible	Num	8	Indicator of "arbitration eligibility"
17	Arb_9192	Num	8	Indicator of "arbitration in 1991/2"
18	Full_Name	Char	19	Players name (in quotation marks)

Then the analysis direction is concentrated on finding the association between salary and other variables.

The means show that the players had moderate to good batting statistics. The range between minimum and maximum of salary is a large value, it indicates great dispersion in the salary (Table 2).

Table 3 . Means for Different Variables

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
salary	Salary (in thousands of dollars)	337	1248.528	1240.013	109.000	6100.000
avg	Batting average	337	0.258	0.040	0.063	0.457
OBP	On-base percentage (OBP)	337	0.324	0.047	0.063	0.486
Runs	Number of runs	337	46.697	29.020	0.000	133.000
Hits	Number of hits	337	92.834	51.896	1.000	216.000
Doubles	Number of doubles	337	16.674	10.452	0.000	49.000
Triples	Number of triples	337	2.338	2.543	0.000	15.000
HRs	Number of home runs	337	9.098	9.290	0.000	44.000
RBI	Number of runs batted in (RBI)	337	44.021	29.559	0.000	133.000
Walks	Number of walks	337	35.018	24.842	0.000	138.000
SOs	Number of strike-outs	337	56.706	33.829	1.000	175.000
SBs	Number of stolen bases	337	8.246	11.665	0.000	76.000
Errors	Number of errors	337	6.772	5.927	0.000	31.000

Understanding the salary is a basis to figure out the next step: the association between salary and other variables. The following steps provide the basic statistical measures, extreme observations and graphs about salary (Table 4, Table 5, Figure 1).

Through attached tables, we learn that some extreme values are equal to 109 or greater than 4000 and the mean is around 1248.528 while median is 740.

Table 4 . Basic Statistical Measures

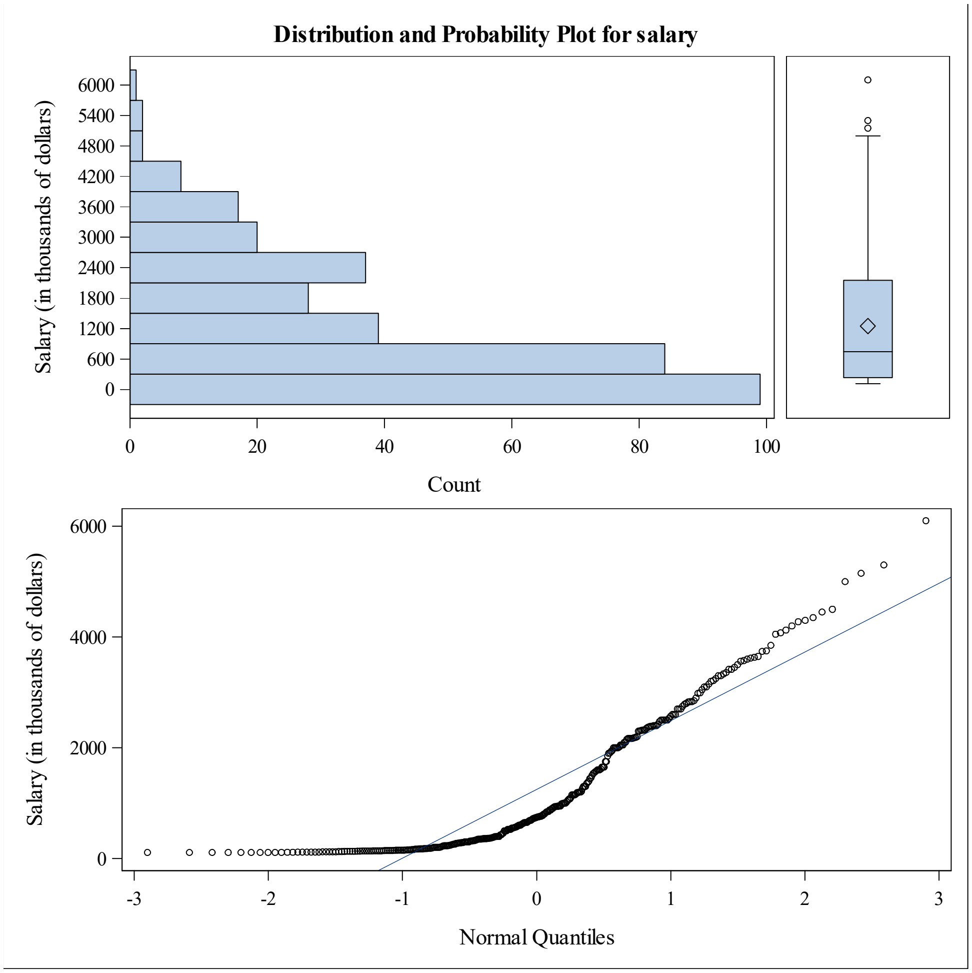
Basic Statistical Measures			
Location		Variability	
Mean	1248.528	Std Deviation	1240
Median	740.000	Variance	1537633
Mode	109.000	Range	5991
		Interquartile Range	1920

Table 5 . Extreme Observations

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
109	337	4500	192
109	322	5000	297
109	284	5150	52
109	268	5300	218
109	230	6100	25

According to the distribution and probability plot for salary, the large proportion of salary of all MLB players was under 2500, and the graph develop an intuitive analysis about basic data of salary.

Figure 1 . Distribution And Probability Plot For Salary



Based on what we learned above, the following steps are try to establish the relationships between salary and other variables.

The number of non-missing values and mean show that the salary of player with eligibility is normally higher than others, and comparing among players who has an eligibility, the salary of player with free agency eligibility is higher than those with arbitration eligibility (Table 5, Figure 2, Figure 3).

Table 5 . Salary With Eligibility (Free Agency And Arbitration)

Analysis Variable : salary Salary (in thousands of dollars)							
Indicator of "free agency eligibility"	Indicator of "arbitration eligibility"	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	0	138	138	262.783	168.106	109.000	935.000
	1	65	65	1567.569	1030.526	109.000	5150.000
1	0	134	134	2108.940	1241.194	109.000	6100.000

Figure 2 . Salary&Free Agency Eligibility

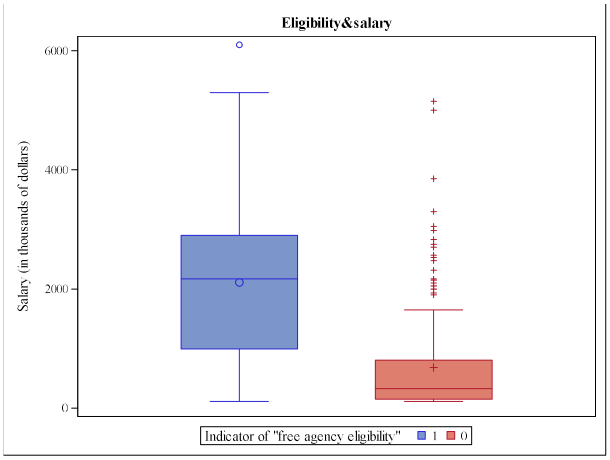
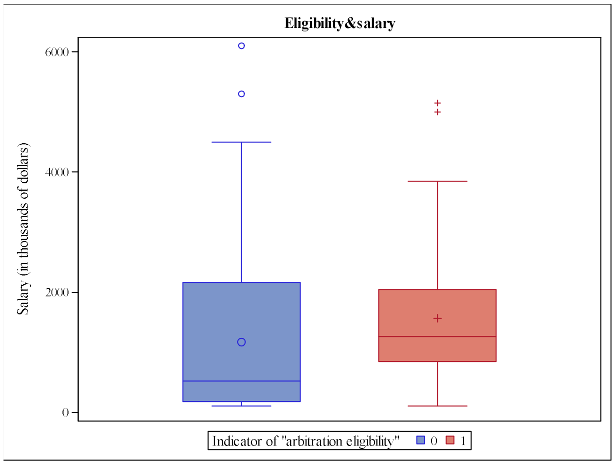


Figure 3 . Salary&Arbitration Eligibility



However, through the percentage of eligibility, we find that, 60.24% of players in free agency have no eligibility, and 80.71% of players in arbitration has no eligibility (Table 6, Table 7). This might be the reason why the mean of salary is under 1250.

Table 6 . Percentage Of Eligibility Of Free Agency

Indicator of "free agency eligibility"				
FA_Eligible	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	203	60.24	203	60.24
1	134	39.76	337	100.00

Table 7 . Percentage Of Eligibility Of Abitration

Indicator of "arbitration eligibility"				
Arb_Eligible	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	272	80.71	272	80.71
1	65	19.29	337	100.00

Next, analysis is continued by exploring the correlation among variables.

The salary has a higher positive correlation with RBI. This means that a high RBI might mean a high of salary (Table 8).

Table 8 . Correlation Between Salary And Other Variables

Pearson Correlation Coefficients, N = 337 Prob > |r| under H0: Rho=0												
	avg	**OBP**	**Runs**	**Hits**	**Doubles**	**Triples**	**HRs**	**RBI**	**Walks**	**SOs**	**SBs**	**Errors**
salary Salary (in thousands of dollars)	0.27637 <.0001	0.32668 <.0001	0.64290 <.0001	0.62124 <.0001	0.57742 <.0001	0.23524 <.0001	0.59045 <.0001	0.66842 <.0001	0.56708 <.0001	0.40549 <.0001	0.25307 <.0001	0.12030 0.0272

The RBI has a higher positive correlation with HRs (Table 9).

Table 9 . Correlation Between RBI And Other Variables

Pearson Correlation Coefficients, N = 337 Prob > |r| under H0: Rho=0							
	Runs	**Hits**	**Doubles**	**Triples**	**HRs**	**Walks**	**SOs**
RBI Number of runs batted in (RBI)	0.83348 <.0001	0.85162 <.0001	0.82537 <.0001	0.33118 <.0001	0.87738 <.0001	0.72706 <.0001	0.74545 <.0001

The correlation between runs and hits is big with 0.923! This again is quite logical because more number of hits means the higher probability to run (Table 10).

Table 10 . Correlation Between Runs And Other Variables

Pearson Correlation Coefficients, N = 337 Prob > |r| under H0: Rho=0									
	avg	**OBP**	**Hits**	**Doubles**	**Triples**	**HRs**	**RBI**	**Walks**	**SBs**
Runs Number of runs	0.43674 <.0001	0.51357 <.0001	0.92317 <.0001	0.83224 <.0001	0.54922 <.0001	0.68106 <.0001	0.83348 <.0001	0.82839 <.0001	0.52612 <.0001

Comment

The interesting part of analysis is because of unfamiliar with baseball dataset, the process of exploring data exists many attempt. With uncertain in the assumption, wrong relation will be drop off by analysis results, and we will be closer to the true answer.

Appendix: SAS Code

SAS Code Group 1 : Input The Data And Label

```
/*Input the data and Label*/
DATA baseball;
   INFILE "D:\GSU\Study\SAS\New folder\BBRawData.txt";
   INPUT
      salary 1-4    avg 6-10          OBP 12-16      Runs 18-20
      Hits 22-24    Doubles 26-27    Triples 29-30  HRs 32-33
      RBI 35-37     Walks 39-41      SOs 43-45      SBs 47-48
      Errors 50-51  FA_Eligible 53   FA_9192 55     Arb_Eligible 57
      Arb_9192 59   Full_Name $61-79;
   LABEL
      Salary='Salary (in thousands of dollars)'
      AVG='Batting average'
      OBP='On-base percentage (OBP)'
      Runs='Number of runs'
      Hits='Number of hits'
      Doubles='Number of doubles'
      Triples='Number of triples'
      HRs='Number of home runs'
      RBI='Number of runs batted in (RBI)'
      Walks='Number of walks'
      SOs='Number of strike-outs'
      SBs='Number of stolen bases'
      Errors='Number of errors'
      FA_Eligible='Indicator of "free agency eligibility"'
      FA_9192='Indicator of "free agent in 1991/2"'
      Arb_Eligible='Indicator of "arbitration eligibility"'
      Arb_9192='Indicator of "arbitration in 1991/2"'
      Full_Name='Players name (in quotation marks)'
      ;
   RUN;
PROC PRINT DATA=baseball;
   RUN;
```

SAS Code Group 2 : Summary Contents About Dataset (Table 1,2)

```
/*To generate summary information about the contents of a dataset*/
PROC CONTENTS DATA=BASEBALL POSITION ;
   TITLE 'baseball dataset structure';
   RUN;
```

SAS Code Group 3 : Means For Different Variables (Table 3)

```
/*To summarize data by central or typical values*/
PROC MEANS DATA=baseball MAXDEC=3;
   VAR salary  avg OBP Runs  Hits Doubles
       Triples HRs RBI Walks SOs  SBs  Errors;
   TITLE 'Mean of each variable';
   RUN;
```

SAS Code Group 4 : Salary Summary Analysis (Table 4,5, Figure 1)

```
/*To get some ideas about salary*/
PROC UNIVARIATE DATA=baseball PLOTS;
   VAR salary;
   TITLE 'salary summary';
   RUN;

PROC SGPLOT DATA=baseball;
   VBOX salary;
   RUN;

PROC SGPLOT DATA=baseball;
   HISTOGRAM salary;
   DENSITY salary;
   RUN;
```

SAS Code Group 5 : Eligibility Influence On Salary (Table 5,6,7, Figure 2,3)

```
/*To find if Eligibility affect players' salaries*/
 PROC MEANS DATA=baseball MAXDEC=3;
    VAR salary;
    CLASS FA_Eligible  Arb_Eligible;
    TITLE 'Eligibility&salary';
    RUN;

/*Percentage among Eligibility*/
PROC FREQ DATA=baseball;
   TABLE FA_Eligible Arb_Eligible;
   RUN;

/*Visualize version of Salary distribution and Eligibility*/
PROC SGPLOT DATA=baseball;
   VBOX salary /GROUP=FA_Eligible;
   RUN;
PROC SGPLOT DATA=baseball;
   VBOX salary /GROUP=Arb_Eligible;
   RUN;
```

SAS Code Group 6 : The Correlation Among Salary And Other Variables (Table 8,9,10)

```
/*To find correlation among variables*/
PROC CORR DATA=baseball PLOTS=matrix;
   VAR avg OBP Runs Hits Doubles Triples HRs RBI Walks SBs ;
   WITH salary;
   TITLE 'Correlation among variables';
   RUN;
PROC CORR DATA=baseball plots=matrix;
   VAR avg OBP Runs Hits Doubles Triples HRs Walks SBs;
   WITH RBI;
   RUN;
PROC CORR DATA=baseball plots=matrix;
   VAR avg OBP  Hits Doubles Triples HRs RBI Walks SBs;
   WITH Runs;
   RUN;
```