An introduction to Proc Means

Joseph Ting

Demographic Analyst

Office of Statistics and Information,
Treasury Board and Finance

April 24th, 2018

Alberta Government

Presentation Outline

- Baseball data – quick overview
- Generating Summary Statistics
- Specifying Statistics and variables
- Grouping observations based on variables
- Variable Formatting
- Creating a SAS dataset out of proc means output (ODS vs. OUTPUT)

Data set

- **Sashelp.baseball**
- **N=322**
- **Various baseball statistics for players at the end of 1982**

	Player's Name	Team at the End of 1986	Home Runs in 1986	League at the End of 1986	Division at the End of 1986	Assists in 1986	1987 Salary in $ Thousands
1	Allanson, Andy	Cleveland	1	American	East	33	.
2	Ashby, Alan	Houston	7	National	West	43	475
3	Davis, Alan	Seattle	18	American	West	82	480
4	Dawson, Andre	Montreal	20	National	East	11	500
5	Galarraga, Andres	Montreal	10	National	East	40	91.5
6	Griffin, Alfredo	Oakland	4	American	West	421	750
7	Newman, Al	Montreal	1	National	East	127	70
8	Salazar, Argenis	Kansas City	0	American	West	283	100
9	Thomas, Andres	Atlanta	6	National	West	290	75
10	Thornton, Andre	Cleveland	17	American	East	0	1100
11	Trammell, Alan	Detroit	21	American	East	445	517.143
12	Trevino, Alex	Los Angeles	4	National	West	45	512.5
13	Van Slyke, Andy	St Louis	13	National	East	11	550

Generating summary statistics

- **Good for exploratory analysis**
- **generates descriptive statistics for continuous variables**
- **Default stats: N, mean, std dev, min, max**
- **Syntax:**

```
proc means data=sashelp.Baseball;
title 'baseball';
run;
```

Results

baseball

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
nAtBat	Times at Bat in 1986	322	390.0745342	143.5958352	127.0000000	687.0000000
nHits	Hits in 1986	322	103.3975155	44.1795091	31.0000000	238.0000000
nHome	Home Runs in 1986	322	11.1024845	8.6987696	0	40.0000000
nRuns	Runs in 1986	322	52.2173913	25.0573661	12.0000000	130.0000000
nRBI	RBIs in 1986	322	49.3726708	25.5011624	8.0000000	121.0000000
nBB	Walks in 1986	322	39.8571429	21.0959408	3.0000000	105.0000000
YrMajor	Years in the Major Leagues	322	7.6801242	4.9697066	1.0000000	24.0000000
CrAtBat	Career Times at Bat	322	2763.08	2328.48	166.0000000	14053.00
CrHits	Career Hits	322	747.6863354	654.7876194	34.0000000	4256.00
CrHome	Career Home Runs	322	74.0900621	90.0651268	0	548.0000000
CrRuns	Career Runs	322	374.2857143	336.4250377	18.0000000	2165.00
CrRbi	Career RBIs	322	347.6149068	338.7903452	9.0000000	1659.00
CrBB	Career Walks	322	273.3944099	273.6253716	8.0000000	1566.00
nOuts	Put Outs in 1986	322	288.9937888	280.6566732	0	1378.00
nAssts	Assists in 1986	322	106.9161491	136.8524541	0	492.0000000
nError	Errors in 1986	322	8.0403727	6.3683591	0	32.0000000
Salary	1987 Salary in $ Thousands	263	535.9258821	451.1186807	67.5000000	2460.00
logSalary	Log Salary	263	5.9272215	0.8891924	4.2121276	7.8079166

Specifying Statistics and Variables

Syntax:

```
proc means data=sashelp.Baseball
        Median Mean RANGE STD SUM;
var nHome nAssts;
title 'Home runs and Assists in 1986'
run;
```

Results

Home runs and Assists in 1986

The MEANS Procedure

Variable	Label	Median	Mean	Range	Std Dev	Sum
nHome	Home Runs in 1986	8.5000000	11.1024845	40.0000000	8.6987696	3575.00
nAssts	Assists in 1986	39.5000000	106.9161491	492.0000000	136.8524541	34427.00

Level of classifications

- **Classifications are used to group your data based on certain variables.**
- **Classification statements discussed are: Ways/ Types/Class/BY**
- **The variables that I will be grouping by right now are: League (American vs. National), Division (East vs. West) and Team.**

Class and Ways Example

- **Question: how does median and mean player salary differ depending on league, division and team?**
- **Use class and ways to create tables, grouping observations based on pairs of these variables**

Results: League and Division

Analysis Variable : Salary 1987 Salary in $ Thousands				
League at the End of 1986	Division at the End of 1986	N Obs	Median	Mean
American	East	85	532.5000000	670.8495588
	West	90	325.0000000	418.5939014
National	East	72	450.0000000	572.3481311
	West	75	450.0000000	487.2592698

Results: Division and Team

The MEANS Procedure

Analysis Variable : Salary 1987 Salary in $ Thousands				
Division at the End of 1986	Team at the End of 1986	N Obs	Median	Mean
East	Baltimore	15	415.0000000	693.5000000
	Boston	10	855.0000000	961.5625000
	Chicago	11	733.3330000	722.4240909
	Cleveland	12	550.0000000	527.7272727
	Detroit	12	420.0000000	497.6190909
	Milwaukee	14	232.5000000	436.2500000
West	Atlanta	11	617.5000000	716.8750000
	California	13	462.5000000	486.4167000
	Chicago	13	185.0000000	354.0909091
	Cincinnati	12	620.0000000	533.8167000
	Houston	11	535.0000000	519.0000000
	Kansas City	14	325.0000000	538.1111111

Results: League and Team

Analysis Variable : Salary 1987 Salary in $ Thousands				
League at the End of 1986	Team at the End of 1986	N Obs	Median	Mean
American	Baltimore	15	415.0000000	693.5000000
	Boston	10	855.0000000	961.5625000
	California	13	462.5000000	486.4167000
	Chicago	13	185.0000000	354.0909091
	Cleveland	12	550.0000000	527.7272727
	Detroit	12	420.0000000	497.6190909
National	Atlanta	11	617.5000000	716.8750000
	Chicago	11	733.3330000	722.4240909
	Cincinnati	12	620.0000000	533.8167000
	Houston	11	535.0000000	519.0000000
	Los Angeles	14	448.7500000	476.0000000
	Montreal	14	250.0000000	359.6111111

Class and Ways Example

Syntax:

```
proc means data=sashelp.Baseball
        Median Mean;
Class League Division Team;
ways 2;
var Salary;
run;
```

Types Example

- **You can create identical output to the last two tables discussed (i.e. League*Team and Division*Team) using the Types statement.**

Syntax:

```
proc means data=sashelp.Baseball
        Median Mean;
Class League Division Team;
types (League Division)*Team;
var Salary;
run;
```

By Statement

- You can add an additional layer of analysis by using the by statement
- Separate output tables are produced for each category in the by-variable
- Must properly sort the data
- Using League as a by-variable

Results: League= American

League at the End of 1986=American

Analysis Variable : Salary 1987 Salary in $ Thousands				
Division at the End of 1986	Team at the End of 1986	N Obs	Median	Mean
East	Baltimore	15	415.0000000	693.5000000
	Boston	10	855.0000000	961.5625000
	Cleveland	12	550.0000000	527.7272727
	Detroit	12	420.0000000	497.6190909
	Milwaukee	14	232.5000000	436.2500000
	New York	11	875.0000000	991.8288889
	Toronto	11	787.5000000	713.6111111
West	California	13	462.5000000	486.4167000
	Chicago	13	185.0000000	354.0909091
	Kansas City	14	325.0000000	538.1111111
	Minneapolis	13	300.0000000	479.2727273
	Oakland	12	512.5000000	539.3750000
	Seattle	12	300.0000000	282.5000000
	Texas	13	235.0000000	311.2272727

Results: League= National

League at the End of 1986=National

Analysis Variable : Salary 1987 Salary in $ Thousands				
Division at the End of 1986	Team at the End of 1986	N Obs	Median	Mean
East	Chicago	11	733.3330000	722.4240909
	Montreal	14	250.0000000	359.6111111
	New York	13	500.0000000	722.5519091
	Philadelphia	12	295.5000000	590.5833000
	Pittsburgh	11	200.0000000	337.1666667
	St Louis	11	550.0000000	621.9697273
West	Atlanta	11	617.5000000	716.8750000
	Cincinnati	12	620.0000000	533.8167000
	Houston	11	535.0000000	519.0000000
	Los Angeles	14	448.7500000	476.0000000
	San Diego	13	333.3335000	506.1805833
	San Francisco	14	195.0000000	276.9230769

By Variable Syntax

```
proc sort data=sashelp.Baseball out=baseball_sorted
By: league;
run;
```

```
proc means data=sashelp.Baseball
        Median Mean;
By: League
Class League Division Team;
var Salary;
run;
```

Problem: Continuous variable as class variable?

Analysis Variable : Salary 1987 Salary in $ Thousands		
Home Runs in 1986	Mean	Median
0	470.5555556	160.0000000
1	223.3333333	225.0000000
2	280.7639167	175.0000000
3	308.5000769	326.6670000
4	393.8725294	277.5000000
5	397.5555556	412.5000000
6	304.8690714	232.5000000
7	372.4615385	300.0000000
8	407.9375000	236.2500000
9	698.0769231	600.0000000
10	540.2333000	550.0000000
11	382.5000000	425.0000000
12	538.1481111	675.0000000
13	648.7727273	500.0000000
14	750.8333333	740.0000000
15	544.3750000	576.2500000
16	425.8333333	200.0000000
17	840.4166667	571.2500000
18	437.6852222	480.0000000
19	987.5000000	990.0000000
20	851.7592222	740.0000000
21	858.9136250	868.7500000
22	571.6666667	750.0000000
23	608.7500000	608.7500000
24	1044.26	1080.73
25	747.5000000	747.5000000
26	802.7776667	850.0000000
27	573.3333333	300.0000000
28	933.1250000	943.7500000
29	819.0000000	535.0000000
30	172.0000000	172.0000000
31	953.6111667	995.8335000
33	190.0000000	190.0000000
34	900.0000000	900.0000000
35	.	.
37	2127.33	2127.33
40	1237.50	1237.50

Solution: Formats

- Use formats to collapse continuous variables for easier analysis

Results

Salary by number of home runs

The MEANS Procedure

Analysis Variable : Salary 1987 Salary in $ Thousands		
Home Runs in 1986	Mean	Median
none	470.5555556	160.0000000
1 to 5	336.8333478	275.0000000
6 to 10	456.2803030	332.5000000
11 to 15	582.1481389	537.5000000
16 to 20	660.2702703	550.0000000
21 to 25	814.5178947	787.5000000
26 or more	833.1234444	850.0000000

Syntax pt.1

```
proc format;
        value nHomefmt
                0='none'
                1-5='1 to 5'
                6-10='6 to 10'
                11-15='11 to 15'
                16-20='16 to 20'
                21-25='21 to 25'
                26-high='26 or more';
run;
```

Syntax pt.2

```
proc means data=sashelp.baseball
        mean median;
class nHome;
var Salary;
format nhome nHomefmt.;
title 'Salary by number of home runs';
run;
```

Save data to a dataset

- You can save the results of the proc means into a SAS dataset if you wish, by using either OUTPUT or ODS
- Output Results are “long”
- ODS Output is “wide”

Output Results

	Home Runs in 1986	_TYPE_	_FREQ_	_STAT_	1987 Salary in $ Thousands
1	.	0	322	N	263
2	.	0	322	MIN	67.5
3	.	0	322	MAX	2460
4	.	0	322	MEAN	535.92588213
5	.	0	322	STD	451.1186807
6	none	1	13	N	9
7	none	1	13	MIN	75
8	none	1	13	MAX	1940
9	none	1	13	MEAN	470.55555556
10	none	1	13	STD	607.94965069
11	1 to 5	1	92	N	69
12	1 to 5	1	92	MIN	67.5
13	1 to 5	1	92	MAX	925
14	1 to 5	1	92	MEAN	336.83334783
15	1 to 5	1	92	STD	246.98799608

ODS Results

	Home Runs in 1986	N Obs	N	Mean	Std Dev	Minimum	Maximum
1	none	13	9	470.55555556	607.94965069	75	1940
2	1 to 5	92	69	336.83334783	246.98799608	67.5	925
3	6 to 10	83	66	456.28030303	350.64596175	70	1600
4	11 to 15	42	36	582.14813889	332.16045378	90	1800
5	16 to 20	41	37	660.27027027	565.51257268	97.5	2460
6	21 to 25	22	19	814.51789474	544.01774388	90	1925.571
7	26 or more	29	27	833.12344444	614.5618086	100	2127.333

Syntax

```
ods output summary=SalaryXnHome_2;
proc means data=sashelp.Baseball;
class nHome;
var Salary;
format nhome nHomefmt.;
title 'Salary by number of home runs';
Output out=SalaryXnHome_1;
run;
```

Thank you!

Questions?