

Summarize Your Data With SAS: Proc Means and Proc Univariate

Nestor Rohowsky, Independent Consultant
LaGrange, IL

Introduction

One hundred men and women are employed by XYZ, Inc. at two regional sites, East and West. The company decides to conduct a survey of employees' salaries. XYZ wants the information summarized:

1. company wide,
2. by gender (male and female),
3. by region (East and West), and
4. by gender within region.

XYZ has available Base SAS® and assigns an analyst the task of obtaining the desired information. How can the analyst use SAS to get the desired results?

SAS Summarization Methods

SAS provides an abundance of options for summarizing data. The choice of which to use depends on how the data are arranged, the programmer's skill and the types of descriptive statistics appropriate to the situation. Note that all summarization methods described below require the data be in numeric format.

1. SAS offers a variety of functions. The basic descriptive statistics, N, Mean, Standard Deviation (STD), Minimum, and Maximum are included as are many lesser used, but no less informative, descriptors such as Coefficient of Variation (CV), Skewness, Kurtosis, Range, Sum, Variance (VAR), Standard Error of the Mean (STDERR), and number of missing observations (NMISS). SAS functions are easiest to use when the data to be summarized are in individual fields. In the case of XYZ, Inc. the data could be arranged as SLRY1, SLRY2, ..., SLRY100. The average salary for the whole company would be calculated as, e.g. `AVGSALRY=MEAN(of SLRY1 - SLRY100)`. Each descriptive statistic requires a separate statement. This method is very inefficient for the task at hand.
2. SAS offers several procedures (PROC), which efficiently do the job of data summarization. These are PROCs MEANS, SUMMARY, TABULATE, and UNIVARIATE. The MEANS, SUMMARY and TABULATE procedures generate the same set of summary statistics. The essential differences among them are that SUMMARY

by default generates its results under NOPRINT conditions, MEANS by default does generate printed output and TABULATE prints output in a table format. PROCs SUMMARY and MEANS are so similar that SAS no longer has a separate chapter in the Procedures manual to describe it. The final procedure, UNIVARIATE generates the same set of summary statistics as the others as well as others not included in any other procedure.

The remainder of this paper will focus on the MEANS and UNIVARIATE procedures.

Salary Survey, Part 1

The XYZ analyst begins the assigned task by sorting and printing the employee salary data. Table 1 displays the salaries printed in descending order, showing employees' numbers, genders and region.

The first attempt at company-wide data summarization is made using PROC MEANS. The program statements are:

```
PROC MEANS DATA=SALARY;
    VAR SALARY;
RUN;
```

These statements generate a default set of summary statistics, N, MEAN, STD DEV, MINIMUM, and MAXIMUM.

The analyst notes that the N=100 employees earn an average of \$101,349.73, ranging from a low of \$8329.00 to a high of \$611,190.00. The standard deviation seems rather large at \$104,180.49.

The analyst decides to apply the same methods to salary data after first sorting by gender, then by region, and finally by region and gender. Output 1 shows all results except those by region and gender.

Salary Survey, Part 2

Close scrutiny of the default results suggest that the default descriptive statistics are not telling the whole story regarding employee salaries. After consulting a basic statistics book and the chapter on PROC MEANS in the SAS Procedures Guide the analyst decides to rerun

the initial analysis by including several optional summary statistics. The revised program statements are

```
PROC MEANS DATA=SALARY N MEAN STD
STDERR MIN MAX RANGE CV SKEWNESS
KURTOSIS;
```

```
VAR SALARY;
```

```
RUN;
```

The requested non-default summary statistics are STDERR (Standard Error of the Mean), CV (Coefficient of Variation), RANGE (Max - Min), SKEWNESS (is the data symmetrically distributed), and KURTOSIS (a measure of how flat or peaked the distribution is). Together, these summary statistics provide a much more complete description of the salary dataset.

Output 2 displays the results company-wide and after sorting by region and gender.

Programming Tip

The above data summaries rely on repeated sorting of the root data set. SAS has an option available in the MEANS and SUMMARY procedures which can make life considerably easier. This is the CLASS statement. Prior data sorting is no longer necessary.

The analyst modifies the prior PROC MEANS call to include the CLASS statement:

```
PROC MEANS DATA=SALARY N MEAN STD
STDERR MIN MAX RANGE CV SKEWNESS
KURTOSIS;
```

```
CLASS GENDER REGION;
```

```
VAR SALARY;
```

```
RUN;
```

The result appears in Output 3. A further modification of the PROC MEANS call to include Both the CLASS and OUTPUT OUT statements as shown below yields Output 4.

```
PROC MEANS DATA=SALARY N MEAN STD
STDERR MIN MAX RANGE CV SKEWNESS
KURTOSIS;
```

```
CLASS GENDER REGION;
```

```
VAR SALARY;
```

```
OUTPUT OUT=SUMSTATS N=N
```

```
MEAN=MEAN STD=STD
```

```
STDERR=STDERR MIN=MIN
```

```
MAX=MAX RANGE=RANGE CV=CV
```

```
SKEWNESS=SKEWNESS
```

```
KURTOSIS=KURTOSIS;
```

```
RUN;
```

```
PROC PRINT DATA=SUMSTATS:
```

```
RUN;
```

This last call provides summary statistics for ALL levels of variables in the CLASS statement and is, therefore, the easiest and most flexible use of the MEANS procedure.

Salary Survey, Part 3

The analyst reviews the different levels of data summary and concludes that the summary information, while extensive, is still incomplete. A return to SAS Procedures Guide leads the analyst to the final, and most comprehensive, summarization procedure, UNIVARIATE.

PROC UNIVARIATE by default generates ALL the descriptive statistics offered as default AND option in PROC MEANS and a host of others including MEDIAN, MODE, 25th and 75th percentiles, the interquartile range, and extreme values. Also offered are several statistical tests with p-values, but these are not of concern to the analyst. Also offered by PROC UNIVARIATE as options are plots of the data. An unfortunate thing about UNIVARIATE is apparently there is no equivalent to the CLASS statement in PROC MEANS. However, the benefits of the UNIVARIATE procedure may outweigh this omission.

The analyst writes new program code as follows:

```
PROC UNIVARIATE DATA=SALARY PLOT;
```

```
VAR SALARY;
```

```
ID EMPLOYEE;
```

```
RUN;
```

Output 5 shows sample results after sorting by gender. The ID EMPLOYEE statement tells the procedure to identify those Male and Female employees with the 5 highest and lowest salaries. The PLOT option yields Stem-and-Leaf, Box and Normal Probability plots. The Stem-and-Leaf plot might be thought of as a histogram on its side. The Box plot shows mean, median, minimum, maximum and 1st and 3rd quartiles of the data. A valuable feature of the Box plots is the identification of unusual data points, perhaps to be interpreted as outliers. The Normal Probability plot visually indicates whether the data are normally distributed (straight line) or not. A final benefit is side-by-side Box plots for the "By" variables.

Discussion

This presentation has compared and contrasted two frequently used summarization procedures offered by SAS, MEANS and UNIVARIATE.

Both are powerful tools for learning about a data set. The first has the benefit of simplicity and compactness; the second offers completeness.

Both procedures permit the data analyst to learn about the data set at hand. Sometimes very simple descriptive statistics will suffice. In other situations more detailed summarization is necessary. Once the data have been properly summarized intelligent decisions can be made about any formal statistical analyses to be performed.

References

SAS Language: Reference, Version 6, 1 ed.,
SAS Institute Inc., Cary, NC, 1993.

SAS Procedures Guide, Version 6, 3 ed., SAS
Institute Inc., Cary, NC, 1990.

® Base SAS is the registered trademark of SAS
Institute Inc., Cary, NC, USA.

For further details please contact the
author at:

Nestor Rohowsky
Statistics and SAS Programming
Consultation
5718 Laurel Avenue
LaGrange, IL 60525
phone: (708) 784-0843
fax: (708) 784-0859
e-mail: MISTERDATA@worldnet.att.net

Table 1
Employee Salaries

EMPLOYEE	REGION	GENDER	SALARY	EMPLOYEE	REGION	GENDER	SALARY
3	East	Male	\$611,190.00	96	East	Female	\$81,032.00
1	East	Female	\$570,580.00	50	West	Male	\$77,463.00
9	West	Male	\$393,271.00	97	East	Female	\$74,748.00
17	West	Female	\$296,865.00	98	West	Female	\$74,185.00
11	East	Male	\$278,374.00	39	East	Male	\$73,986.00
10	East	Female	\$254,312.00	56	West	Male	\$73,695.00
27	East	Female	\$244,080.00	100	West	Female	\$72,901.00
20	West	Male	\$224,911.00	72	West	Female	\$69,577.00
4	East	Female	\$222,400.00	76	West	Female	\$67,321.00
28	West	Male	\$214,187.00	67	West	Female	\$67,025.00
37	East	Male	\$208,434.00	58	West	Male	\$65,115.00
26	East	Female	\$192,296.00	73	East	Female	\$63,392.00
30	East	Male	\$182,230.00	85	West	Female	\$61,649.00
47	East	Male	\$181,422.00	70	West	Male	\$60,403.00
29	West	Male	\$171,015.00	92	East	Male	\$58,774.00
81	East	Male	\$162,566.00	2	West	Male	\$44,483.00
75	West	Male	\$159,191.00	5	West	Male	\$28,935.00
83	East	Female	\$158,856.00	6	West	Female	\$25,421.00
12	East	Female	\$155,880.00	7	West	Male	\$25,223.00
8	East	Female	\$153,536.00	13	East	Male	\$17,666.00
25	West	Male	\$153,351.00	16	West	Female	\$16,829.00
18	West	Female	\$152,965.00	15	East	Male	\$16,594.00
60	East	Male	\$150,978.00	19	East	Female	\$14,260.00
91	East	Female	\$147,232.00	22	East	Female	\$13,932.00
38	East	Female	\$142,840.00	21	West	Male	\$13,743.00
46	West	Female	\$141,941.00	33	East	Male	\$11,602.00
24	East	Female	\$136,308.00	34	East	Female	\$11,540.00
86	East	Male	\$133,770.00	41	East	Male	\$11,362.00
63	East	Female	\$121,352.00	35	East	Female	\$11,196.00
45	East	Female	\$119,472.00	44	East	Male	\$11,158.00
74	East	Male	\$117,158.00	49	West	Female	\$10,825.00
14	East	Female	\$116,484.00	36	West	Male	\$10,811.00
48	East	Female	\$115,268.00	40	West	Male	\$10,479.00
43	West	Male	\$114,243.00	53	West	Female	\$10,141.00
93	East	Male	\$113,222.00	61	West	Female	\$10,009.00
66	West	Female	\$112,549.00	68	East	Female	\$9,788.00
62	West	Female	\$111,513.00	55	West	Male	\$9,651.00
59	East	Female	\$110,928.00	69	East	Male	\$9,506.00
51	East	Male	\$108,518.00	57	West	Male	\$9,367.00
65	West	Female	\$107,097.00	64	West	Female	\$9,269.00
79	West	Female	\$106,853.00	78	East	Female	\$9,080.00
23	West	Female	\$103,101.00	80	East	Male	\$8,954.00
82	West	Male	\$102,383.00	90	East	Female	\$8,868.00
32	West	Male	\$102,211.00	94	West	Female	\$8,789.00
99	East	Female	\$99,028.00	95	East	Male	\$8,642.00
52	West	Female	\$98,485.00	71	East	Male	\$8,546.00
77	West	Female	\$98,209.00	89	West	Female	\$8,521.00
84	West	Female	\$98,149.00	88	East	Female	\$8,452.00
42	East	Male	\$92,146.00	87	West	Female	\$8,329.00
54	West	Male	\$89,607.00				
31	West	Male	\$88,979.00				

Output 1

Default PROC MEANS Output

Analysis Variable : SALARY

N	Mean	Std Dev	Minimum	Maximum
100	101349.73	104180.49	8329.00	611190.00

----- GENDER=Female -----

N	Mean	Std Dev	Minimum	Maximum
54	98434.41	96267.15	8329.00	570580.00

----- GENDER=Male -----

N	Mean	Std Dev	Minimum	Maximum
46	104772.07	113757.62	8546.00	611190.00

----- REGION=East -----

N	Mean	Std Dev	Minimum	Maximum
51	116543.88	123056.12	8452.00	611190.00

----- REGION=West -----

N	Mean	Std Dev	Minimum	Maximum
49	85535.41	78173.80	8329.00	393271.00

Output 2

Enhanced PROC MEANS Output

Analysis Variable : SALARY

N	Mean	Std Dev	Std Error	Minimum	Maximum	Range	CV	Skewness	Kurtosis
100	101349.73	104180.49	10418.06	8329.00	611190.00	602861.00	102.7930697	2.4257881	8.7374036

REGION=East GENDER=Female

N	Mean	Std Dev	Std Error	Minimum	Maximum	Range	CV	Skewness	Kurtosis
28	120247.88	115804.81	21828.36	8452.00	570880.00	562428.00	96.0556059	2.2370180	7.8107189

REGION=East GENDER=Male

N	Mean	Std Dev	Std Error	Minimum	Maximum	Range	CV	Skewness	Kurtosis
23	112034.70	134176.28	27877.68	8548.00	611190.00	602642.00	119.7631327	2.5014029	8.3243943

REGION=West GENDER=Female

N	Mean	Std Dev	Std Error	Minimum	Maximum	Range	CV	Skewness	Kurtosis
25	74843.00	84275.18	12805.40	8329.00	298885.00	288556.00	85.7854189	1.5733775	4.4821754

REGION=West GENDER=Male

N	Mean	Std Dev	Std Error	Minimum	Maximum	Range	CV	Skewness	Kurtosis
23	97509.43	91412.54	19060.63	9367.00	393271.00	383904.00	93.7473781	1.7023345	3.8132755

Output 3

Descriptive Statistics Using the "CLASS" Statement
Default Output

Analysis Variable : SALARY

REGION	GENDER	N Obs	N	Mean	Std Dev	Std Error	Minimum	Maximum	Range	CV	Skewness	Kurtosis
1	1	28	28	120247.86	115504.81	21828.36	8452.00	570580.00	562128.00	96.0556059	2.2370180	7.8107169
	2	23	23	112034.70	134176.26	27977.68	8546.00	611190.00	602644.00	119.7631327	2.5014029	8.3243943
2	1	26	26	74943.00	84275.18	12805.40	8329.00	296865.00	288538.00	85.7654199	1.5733775	4.4621754
	2	23	23	97509.43	91412.54	19080.83	9367.00	393271.00	383904.00	93.7473781	1.7023345	3.8132755

Output 4

Descriptive Statistics Using the "CLASS" Statement
Print of "OUTPUT" Data Set

OBS	REGION	GENDER	_TYPE_	_FREQ_	N	MEAN	STD	STDERR	MIN	MAX	RANGE	CV	SKENNESS	KURTOSIS
1	.	.	0	100	100	101349.73	104180.49	10418.05	8329	611190	602861	102.793	2.42579	8.7374
2	.	Female	1	54	54	98434.41	96267.15	13100.30	8329	570580	562251	97.798	2.46672	10.0216
3	.	Male	1	46	46	104772.07	113757.62	16772.65	8546	611190	602644	108.576	2.40067	8.2177
4	East	.	2	51	51	116543.88	123056.12	17231.30	8452	611190	602738	105.588	2.30621	7.2438
5	West	.	2	49	49	85535.41	78173.80	11167.69	8329	393271	384942	91.393	1.77730	4.5189
6	East	Female	3	28	28	120247.86	115504.81	21828.36	8452	570580	562128	96.058	2.23702	7.8107
7	East	Male	3	23	23	112034.70	134176.26	27977.68	8546	611190	602644	119.763	2.50140	8.3244
8	West	Female	3	26	26	74943.00	84275.18	12805.40	8329	296865	288538	85.765	1.57338	4.4622
9	West	Male	3	23	23	97509.43	91412.54	19080.83	9367	393271	383904	93.747	1.70233	3.8133

Output 5

Descriptive Statistics Using PROC UNIVARIATE - Subset by Gender

GENDER=Female

Univariate Procedure

le=SALARY

Moments

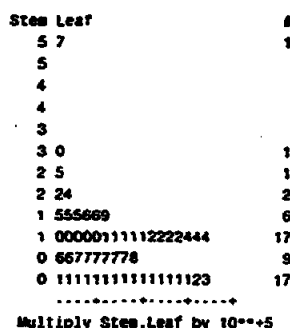
N 54
Mean 98434.41
Std Dev 96267.15
Skewness 2.466725
Kurtosis 10.0216
CSS 1.014E12
Std Mean 13100.3
Pr>|T| 0.0001
Num > 0 54
Pr>=|M| 0.0001
Sgn Rank 742.5
Pr>=|S| 0.0001

Quantiles(Def=5)

100% Max 570580
75% Q3 136308
50% Med 98179
25% Q1 13932
0% Min 8329
Range 562251
Q3-Q1 122376
Mode 8329

Extremes

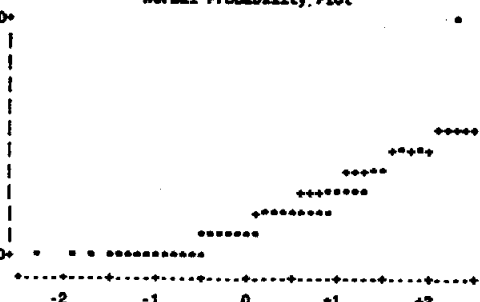
Lowest ID Highest ID
8329(87) 222400(4)
8452(88) 244080(27)
8521(89) 254312(10)
8789(94) 296885(17)
8888(90) 570580(1)



Boxplot



Normal Probability Plot



GENDER=Male

Univariate Procedure

Variable=SALARY

Moments

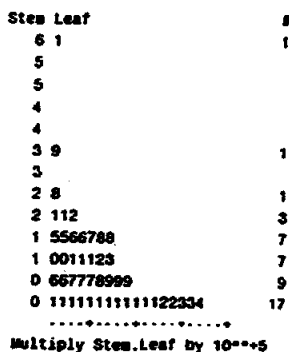
N 46
Mean 104772.1
Std Dev 113757.6
Skewness 2.400668
Kurtosis 8.217695
CSS 1.087E12
Std Mean 16772.65
Pr>|T| 0.0001
Num > 0 46
Pr>=|M| 0.0001
Sgn Rank 540.5
Pr>=|S| 0.0001

Quantiles(Def=5)

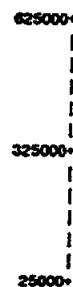
100% Max 611190
75% Q3 153351
50% Med 83221
25% Q1 13743
0% Min 8546
Range 602644
Q3-Q1 139608
Mode 8546

Extremes

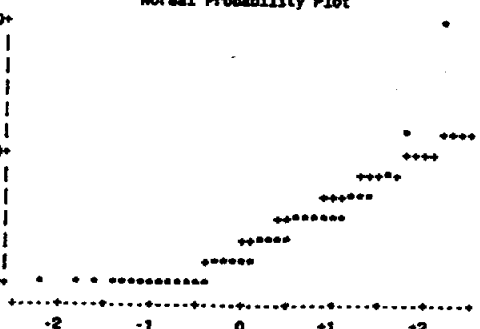
Lowest ID Highest ID
8546(71) 214187(28)
8642(95) 224911(20)
8854(80) 278374(11)
9367(57) 393271(9)
9506(69) 611190(3)



Boxplot



Normal Probability Plot



Output 5

Descriptive Statistics Using PROC UNIVARIATE - Subset by Gender

Univariate Procedure
Schematic Plots

