STAT604 SAS Lesson 16

SAS

Producing Descriptive Statistics

The FREQ Procedure – Prep Guide Chapter 15

SAS

Business Scenario

A new data set, **orion.nonsales2**, must be validated. It contains information about non-sales employees and might include invalid and missing values.

Partial **orion.nonsales2**

Employee_ID	First	Last	Gender	Salary	Job_Title	Country
120101	Patrick	Lu	M	163040	Director	AU
120104	Kareen	Billington	F	46230	Admin Mgr	au
120105	Liz	Povey	F	27110	Secretary I	AU
120106	John	Hornsey	M	.	Office Asst II	AU
120107	Sherie	Sheedy	F	30475	Office Asst II	AU
120108	Gladys	Gromek	F	27660	Warehouse Asst II	AU

Considerations

Use the FREQ procedure to screen for invalid, missing, and duplicate data values.

Requirements of non-sales employee data:

- **Employee_ID** values must be unique and not missing.

- **Gender** must be *F* or *M*.

- **Job_Title** must not be missing.

- **Country** must have a value of *AU* or *US*.

- **Salary** values must be in the numeric range of 24000 to 500000.

Short Answer Poll

What problems exist with the data in this partial data set?

Employee_ID	First	Last	Gender	Salary	Job_Title	Country
120101	Patrick	Lu	M	163040	Director	AU
120104	Kareen	Billington	F	46230	Administration Manager	au
120105	Liz	Povey	F	27110	Secretary I	AU
120106	John	Hornsey	M	.	Office Assistant II	AU
120107	Sherie	Sheedy	F	30475	Office Assistant III	AU
120108	Gladys	Gromek	F	27660	Warehouse Assistant II	AU
120108	Gabriele	Baker	F	26495	Warehouse Assistant I	AU
120110	Dennis	Entwisle	M	28615	Warehouse Assistant III	AU
120111	Ubaldo	Spillane	M	26895	Security Guard II	AU
120112	Ellis	Glattback	F	26550		AU
120113	Riu	Horsey	F	26870	Security Guard II	AU
120114	Jeannette	Buddery	G	31285	Security Manager	AU
120115	Hugh	Nichollas	M	2650	Service Assistant I	AU
.	Austen	Ralston	M	29250	Service Assistant II	AU
120117	Bill	Mccleary	M	31670	Cabinet Maker III	AU
120118	Darshi	Hartshorn	M	28090	Cabinet Maker II	AU

Hint: There are seven data problems.

§sas

Short Answer Poll – Correct Answer

What problems exist with the data in this partial data set?

Employee_ID	First	Last	Gender	Salary	Job_Title	Country
120101	Patrick	Lu	M	163040	Director	AU
120104	Kareen	Billington	F	46230	Administration Manager	au
120105	Liz	Povey	F	27110	Secretary I	AU
120106	John	Hornsey	M	.	Office Assistant II	AU
120107	Sherie	Sheedy	F	30475	Office Assistant III	AU
120108	Gladys	Gromek	F	27660	Warehouse Assistant II	AU
120108	Gabriele	Baker	F	26495	Warehouse Assistant I	AU
120110	Dennis	Entwisle	M	28615	Warehouse Assistant III	AU
120111	Ubaldo	Spillane	M	26895	Security Guard II	AU
120112	Ellis	Glattback	F	26550		AU
120113	Riu	Horsey	F	26870	Security Guard II	AU
120114	Jeannette	Buddery	G	31285	Security Manager	AU
120115	Hugh	Nichollas	M	2650	Service Assistant I	AU
.	Austen	Ralston	M	29250	Service Assistant II	AU
120117	Bill	Mccleary	M	31670	Cabinet Maker III	AU
120118	Darshi	Hartshorn	M	28090	Cabinet Maker II	AU

Hint: There are seven data problems.

§.sas

FREQ Procedure for Data Validation

The FREQ procedure lists all discrete values for a variable and reports missing values.

```
proc freq data=orion.nonsales2;
    tables Gender Country / nocum nopercent;
run;
```

Viewing the Output

PROC FREQ Output

```
The FREQ Procedure

Gender      Frequency
_____

F                110
G                  1
M                123

Frequency Missing = 1


Country     Frequency
_____

AU                33
US               196
au                 3
us                 3
```

§sas

NLEVELS Option

The *NLEVELS option* displays a table that provides the number of distinct values for each analysis variable.

```
proc freq data=orion.nonsales2 nlevels;
    tables Gender Country / nocum nopercent;
run;
```

PROC FREQ DATA=*SAS-data-set* NLEVELS;
    TABLES *variable(s)* ;
RUN;

Viewing the Output

PROC FREQ Output

Number of Variable Levels			
Variable	**Levels**	**Missing Levels**	**Nonmissing Levels**
Gender	4	1	3
Country	4	0	4

Gender	Frequency
F	110
G	1
M	123
Frequency Missing = 1	

Country	Frequency
AU	33
US	196
au	3
us	3

§sas

Check for Uniqueness

The values of **Employee_ID** must be unique and not missing. PROC FREQ can be used to check for duplicate or missing **values**.

```
proc freq data=orion.nonsales2 order=freq;
    tables Employee_ID / nocum nopercent;
run;
```

The ORDER=FREQ option displays the results in descending frequency order.

Viewing the Output

Partial PROC FREQ Output

```
The FREQ Procedure

Employee_ID      Frequency
_____

120108                   2
120101                   1
120104                   1
120105                   1
120106                   1
```

```
121134                1
121141                1
121142                1
121146                1
121147                1
121148                1

Frequency Missing = 1
```

NLEVELS Option

NLEVELS can also be used to identify duplicates, when the number of distinct values is known.

```
proc freq data=orion.nonsales2 nlevels;
    tables Employee_ID / noprint;
run;
```

This example uses the NOPRINT option to suppress the frequency table. Only the Number of Variable Levels table is displayed.

Viewing the Output

[handwritten: → Just guzz vs the top table from slide 10]

Partial PROC FREQ Output

```
                    The FREQ Procedure

                Number of Variable Levels

                                Missing      Nonmissing
Variable            Levels       Levels          Levels
Employee_ID            234            1             233
```

There are 235 employees, but there are only 234 distinct **Employee_ID**
values. Therefore, there is one duplicate value and one missing value for
Employee_ID.

NLEVELS Option

The _ALL_ keyword with the NOPRINT option displays the number of levels for all variables without displaying frequency counts.

```
proc freq data=orion.nonsales2 nlevels;
    tables _all_ / noprint;
run;
```

Viewing the Output

PROC FREQ Output

```
                    The FREQ Procedure
                 Number of Variable Levels

                                Missing      Nonmissing
    Variable          Levels     Levels          Levels
    _____

    Employee_ID          234          1             233
    First                204          0             204
    Last                 228          0             228
    Gender                 4          1               3
    Salary               230          1             229
    Job_Title            125          1             124
    Country                4          0               4
```

No frequency tables were displayed.

§sas

Identifying Observations with Invalid Data

PROC FREQ uncovered the existence of invalid data values for **Gender**, **Country**, and **Employee_ID**. Use PROC PRINT to display the observations with invalid values.

```
proc print data=orion.nonsales2;
   where Gender not in ('F','M') or
         Country not in ('AU','US') or
            Job_Title is null or
            Employee_ID is missing or
            Employee_ID=120108;
run;
```

§sas

Viewing the Output

PROC PRINT Output

```
Obs    Employee_ID  First       Last        Gender   Salary  Job_Title                 Country

  2        120104    Kareen      Billington    F       46230  Administration Manager      au
  6        120108    Gladys      Gromek        F       27660  Warehouse Assistant II      AU
  7        120108    Gabriele    Baker         F       26495  Warehouse Assistant I       AU
 10        120112    Ellis       Glattback     F       26550                              AU
 12        120114    Jeannette   Buddery       G       31285  Security Manager            AU
 14             .    Austen      Ralston       M       29250  Service Assistant II        AU
 84        120695    Trent       Moffat        M       28180  Warehouse Assistant II      au
 87        120698    Geoff       Kistanna      M       26160  Warehouse Assistant I       au
101        120723    Deanna      Olsen                 33950  Corp. Comm. Specialist II   US
125        120747    Zashia      Farthing      F       43590  Financial Controller I      us
197        120994    Danelle     Sergeant      F       31645  Office Administrator I      us
200        120997    Mary        Donathan      F       27420  Shipping Administrator I    us
```
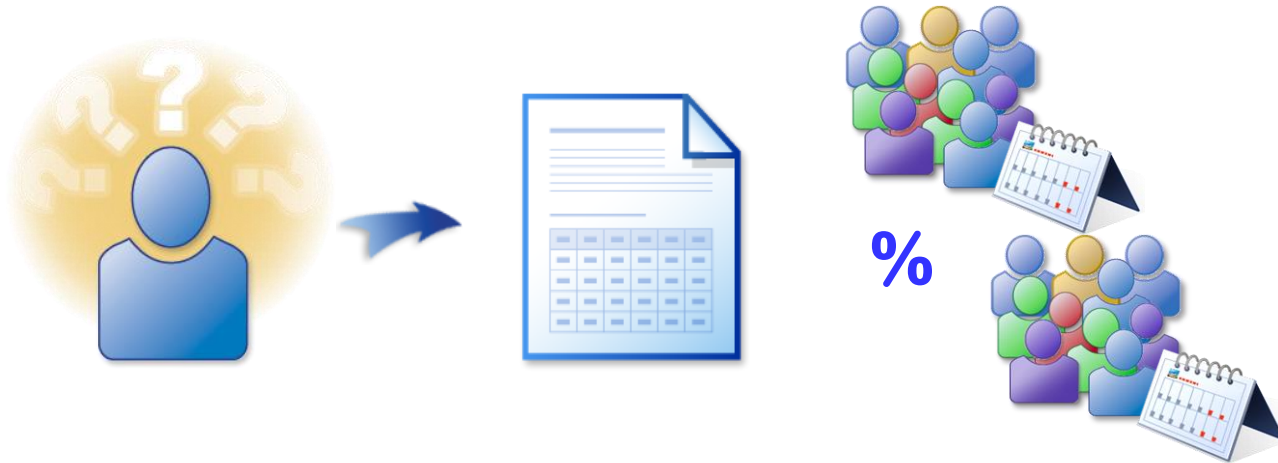
original observation numbers

§sas

Business Scenario

The manager of Human Resources requested a report that shows the number and percent of sales employees who are hired each year.

Using Formats in PROC FREQ

A FORMAT statement can be used in PROC FREQ to format data values.

```
proc freq data=orion.sales;
    tables Hire_Date / nocum;
    format Hire_Date date9.;
run;
```

Partial PROC FREQ Output

```
                The FREQ Procedure


Hire_Date        Frequency          Percent

01JAN1978              17            10.30
01FEB1978               2             1.21
01APR1978               1             0.61
01JUL1978               1             0.61
01AUG1978               1             0.61
```

many discrete values, and not what the manager requested

§sas

Using Formats in PROC FREQ

A FORMAT statement can also be used in PROC FREQ to group the data.

```
proc freq data=orion.sales;
    tables Hire_Date / nocum;
    format Hire_Date year4.;
run;
```

Partial PROC FREQ Output

```
              The FREQ Procedure


    Hire_Date     Frequency         Percent

         1978            23           13.94
         1979             2            1.21
         1980             4            2.42
         1981             3            1.82
         1982             7            4.24
```

fewer discrete values

Short Answer Poll

Can user-defined formats be used to group data?

Short Answer Poll – Correct Answer

Can user-defined formats be used to group data? **yes**

Salary	Frequency	Percent	Cumulative Frequency	Cumulative Percent
		The FREQ Procedure		
Tier1	1	0.61	1	0.61
Tier2	158	95.76	159	96.36
Tier3	4	2.42	163	98.79
Tier4	2	1.21	165	100.00

Producing Descriptive Statistics

The MEANS Procedure – Prep Guide Chapter 15

Business Scenario

The payroll manager would like to see the average salary for all employees.
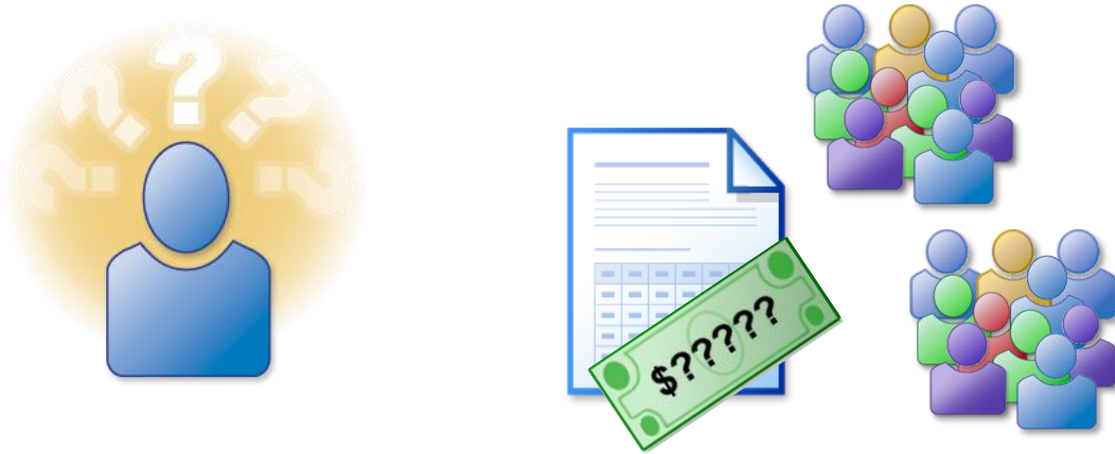
MEANS Procedure

The MEANS procedure produces summary reports with descriptive statistics.

```
proc means data=orion.sales;
run;
```

PROC MEANS DATA=*input-data-set <options statistics>*;
    <VAR *analysis-variable(s)*;>
    <CLASS *classification-variable(s)*;>
RUN;

- *Analysis variables* are the **numeric** variables for which statistics are to be computed.

- *Classification variables* are variables whose values define subgroups for the analysis. They can be character or numeric.
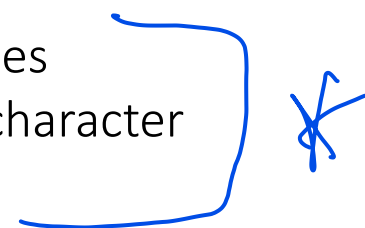
Viewing the Output
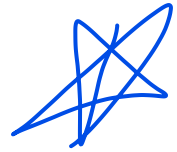
PROC MEANS Output

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Employee_ID	165	120713.90	450.0866939	120102.00	121145.00
Salary	165	31160.12	20082.67	22710.00	243190.00
Birth_Date	165	3622.58	5456.29	-5842.00	10490.00
Hire_Date	165	12054.28	4619.94	5114.00	17167.00

Default statistics are displayed for all numeric variables.

VAR Statement

The VAR statement identifies the analysis variable (or variables) and their order in the output.

```
proc means data=orion.sales;
   var Salary;
run;
```
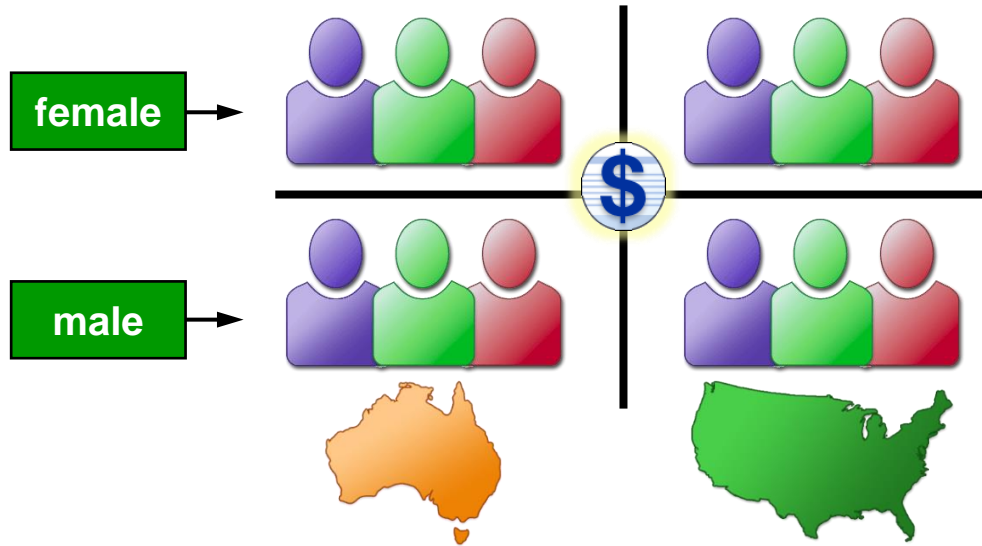
VAR *variable(s)*;

The MEANS Procedure

	Analysis Variable : Salary			
N	Mean	Std Dev	Minimum	Maximum
165	31160.12	20082.67	22710.00	243190.00

Business Scenario

Analyze **Salary** by **Country** within **Gender**.

CLASS Statement

The *CLASS statement* identifies variables whose values define subgroups for the analysis.

```
proc means data=orion.sales;
   var Salary;
   class Gender Country;
run;
```

CLASS *classification-variable(s)*;

- Classification variables are character or numeric.
- They typically have few discrete values.
- The data set does **not** need to be sorted or indexed by the classification variables.

Viewing the Output

Statistics are produced for each combination of values of the classification variables.

The MEANS Procedure

Analysis Variable : Salary							
Gender	Country	N Obs	N	Mean	Std Dev	Minimum	Maximum
F	AU	27	27	27702.41	1728.23	25185.00	30890.00
	US	41	41	29460.98	8847.03	25390.00	83505.00
M	AU	36	36	32001.39	16592.45	25745.00	108255.00
	US	61	61	33336.15	29592.69	22710.00	243190.00

- *N Obs* – the number of observations with each unique combination of **class** variables

- *N* – the number of observations with nonmissing values of the analysis variable (or variables)

Short Answer Poll

For a given data set, there are 63 observations with a **Country** value of *AU*. Of those 63 observations, only 61 observations have a value for **Salary**.

Which output is correct?

a.

```
Analysis Variable : Salary

                    N
    Country        Obs           N
   _____

    AU             63           61
   _____
```

b.

```
Analysis Variable : Salary

                    N
    Country        Obs           N
   _____

    AU             61           63
   _____
```

Short Answer Poll – Correct Answer

For a given data set, there are 63 observations with a **Country** value of *AU*. Of those 63 observations, only 61 observations have a value for **Salary**.

Which output is correct?

a.

```
Analysis Variable : Salary

                    N
Country      Obs           N
_____

AU              63          61
_____
```
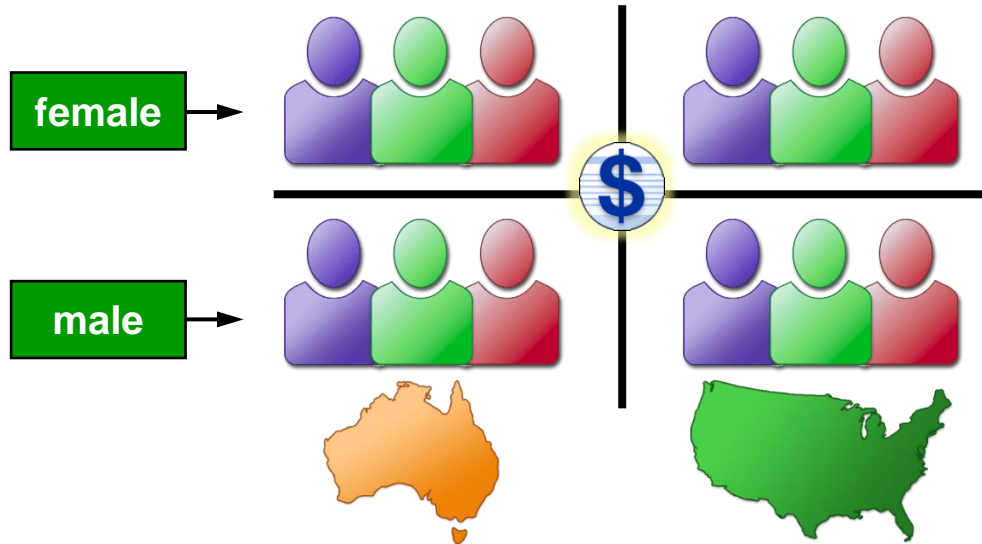
b.

```
Analysis Variable : Salary

                    N
Country      Obs           N
_____

AU              61          63
_____
```

Business Scenario

Analyze **Salary** by **Country** within **Gender**. Generate a report that includes the number of missing **Salary** values, as well as the minimum, maximum, and sum of salaries.

PROC MEANS Statistics

Use options in the PROC MEANS statement to request specific statistics. ☆

```
proc means data=orion.sales nmiss min max sum;
    var Salary;
    class Gender Country;
run;
```

(handwritten note) ☆ once we specify I must specify all the ones we want

The requested statistics override the default statistics.

PROC MEANS Statistics

The statistics are displayed in the order in which they are requested.

The MEANS Procedure

	Analysis Variable : Salary					
Gender	Country	N Obs	N Miss	Minimum	Maximum	Sum
F	AU	27	0	25185.00	30890.00	747965.00
	US	41	0	25390.00	83505.00	1207900.00
M	AU	36	0	25745.00	108255.00	1152050.00
	US	61	0	22710.00	243190.00	2033505.00

Other PROC MEANS Statistics

mode not available in R.

Descriptive Statistic Keywords

CLM	CSS	CV	LCLM	MAX
MEAN	MIN	MODE	N	NMISS
KURTOSIS	RANGE	SKEWNESS	STDDEV	STDERR
SUM	SUMWGT	UCLM	USS	VAR

Quantile Statistic Keywords

MEDIAN | P50	P1	P5	P10	Q1 | P25
Q3 | P75	P90	P95	P99	QRANGE

Hypothesis Testing Keywords

PROBT	T			

§sas

PROC MEANS Statement Options

Options can also be placed in the PROC MEANS statement.

Option	Description
MAXDEC=	Specifies the number of decimal places to display.
NONOBS	Suppresses the N Obs column.

§sas

MAXDEC= Option

MAXDEC=0

The MEANS Procedure

Analysis Variable : Salary						
Country	N Obs	N	Mean	Std Dev	Minimum	Maximum
AU	63	63	30159	12699	25185	108255
US	102	102	31778	23556	22710	243190

MAXDEC=1

The MEANS Procedure

Analysis Variable : Salary						
Country	N Obs	N	Mean	Std Dev	Minimum	Maximum
AU	63	63	30159.0	12699.1	25185.0	108255.0
US	102	102	31778.5	23555.8	22710.0	243190.0

NONOBS Option

N Obs included by default

Analysis Variable : Salary

Country	N Obs	N	Mean	Std Dev	Minimum	Maximum
AU	63	63	30158.97	12699.14	25185.00	108255.00
US	102	102	31778.48	23555.84	22710.00	243190.00

NONOBS option

Analysis Variable : Salary

Country	N	Mean	Std Dev	Minimum	Maximum
AU	63	30158.97	12699.14	25185.00	108255.00
US	102	31778.48	23555.84	22710.00	243190.00

BY Statement

The BY statement is used to request separate analyses (tables) for each BY group.

```
proc means data=sorted;
   var salary;
   class Gender;
   by Country;
run;
```

The data set must be sorted or indexed by the variable (or variables) named in the BY statement.

§sas

BY Statement

Country=AU

			Analysis Variable : Salary			
Gender	N Obs	N	Mean	Std Dev	Minimum	Maximum
F	27	27	27702.41	1728.23	25185.00	30890.00
M	36	36	32001.39	16592.45	25745.00	108255.00

Country=US

			Analysis Variable : Salary			
Gender	N Obs	N	Mean	Std Dev	Minimum	Maximum
F	41	41	29460.98	8847.03	25390.00	83505.00
M	61	61	33336.15	29592.69	22710.00	243190.00

Output Data Sets

PROC MEANS produces output data sets using the following method:

> **OUTPUT OUT=**_SASdataset <options>_**;**

The output data set contains the following variables:

- BY variables

- class variables

- the automatic variables **_TYPE_** and **_FREQ_**

- the variables requested in the OUTPUT statement or **_STAT_** if you do not specify variables for statistics

OUTPUT Statement OUT= Option

The statistics in the PROC statement impact only the MEANS report, not the data set.

```
proc means data=orion.sales sum mean range;
    var Salary;
    class Gender Country;
    output out=work.means1;
run;

proc print data=work.means1;
run;
```

OUTPUT Statement OUT= Option

Partial PROC PRINT Output

Obs	Gender	Country	_TYPE_	_FREQ_	_STAT_	Salary
1			0	165	N	165.00
2			0	165	MIN	22710.00
3			0	165	MAX	243190.00
4			0	165	MEAN	31160.12
5			0	165	STD	20082.67
6		AU	1	63	N	63.00
7					MIN	25185.00
8					MAX	108255.00
9		AU	1	63	MEAN	30158.97
10		AU	1	63	STD	12699.14
11		US	1	102	N	102.00
12		US	1	102	MIN	22710.00
13		US	1	102	MAX	243190.00
14		US	1	102	MEAN	31778.48
15		US	1	102	STD	23555.84
16	F		2	68	N	68.00
17	F		2	68	MIN	25185.00
18	F		2	68	MAX	83505.00
19	F		2	68	MEAN	28762.72
20	F		2	68	STD	6974.15

default statistics

OUTPUT Statement OUT= Option

The OUTPUT statement can also do the following:

- specify the statistics for the output data set

- select and name variables

```
proc means data=orion.sales noprint;
   var Salary;
   class Gender Country;
   output out=work.means2
           min=minSalary max=maxSalary
           sum=sumSalary mean=aveSalary;
run;

proc print data=work.means2;
run;
```

The NOPRINT option suppresses the display of all output.

OUTPUT Statement OUT= Option

PROC PRINT Output

Obs	Gender	Country	_TYPE_	_FREQ_	min Salary	max Salary	sum Salary	ave Salary
1			0	165	22710	243190	5141420	31160.12
2		AU	1	63	25185	108255	1900015	30158.97
3		US	1	102	22710	243190	3241405	31778.48
4	F		2	68	25185	83505	1955865	28762.72
5	M		2	97	22710	243190	3185555	32840.77
6	F	AU	3	27	25185	30890	747965	27702.41
7	F	US	3	41	25390	83505	1207900	29460.98
8	M	AU	3	36	25745	108255	1152050	32001.39
9	M	US	3	61	22710	243190	2033505	33336.15

OUTPUT Statement OUT= Option

`_TYPE_` is a numeric variable that shows which combination of class variables produced the summary statistics in that observation.

PROC PRINT Output

OUTPUT Statement OUT= Option

Obs	Gender	Country	_TYPE_	_FREQ_	min Salary	max Salary	sum Salary	ave Salary
1			0	165	22710	243190	5141420	31160.12
2		AU	1	63	25185	108255	1900015	30158.97
3		US	1	102	22710	243190	3241405	31778.48
4	F		2	68	25185	83505	1955865	28762.72
5	M		2	97	22710	243190	3185555	32840.77
6	F	AU	3	27	25185	30890	747965	27702.41
7	F	US	3	41	25390	83505	1207900	29460.98
8	M	AU	3	36	25745	108255	1152050	32001.39
9	M	US	3	61	22710	243190	2033505	33336.15

TYPE	Type of Summary	_FREQ_
0	overall summary	165
1	summary by `Country` only	63 AU + 102 US = 165
2	summary by `Gender` only	68 F + 97 M = 165
3	summary by `Country` and `Gender`	27 F AU + 41 F US + 36 M AU + 61 M US = 165

§sas

Idea Exchange

Which PROC MEANS statistics would you request when you are validating numeric variables?

min/max

~really doesn't help w/ validating.
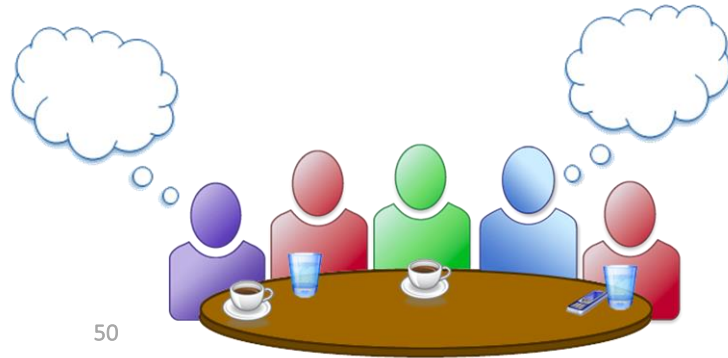
Producing Descriptive Statistics

The UNIVARIATE Procedure

↳ use for extreme values.

Business Scenario

Validate salary data in **orion.nonsales2**. **Salary** must be in the numeric range of 24000 to 500000.

Partial **orion.nonsales2**

Employee _ID	First	Last	Gender	Salary	Job_Title	Country
120101	Patrick	Lu	M	163040	Director	AU
120104	Kareen	Billington	F	46230	Admin Mgr	au
120105	Liz	Povey	F	27110	Secretary I	AU
120106	John	Hornsey	M	.	Office Asst II	AU
120107	Sherie	Sheedy	F	30475	Office Asst II	AU
120108	Gladys	Gromek	F	27660	Warehouse Asst II	AU

UNIVARIATE Procedure

PROC UNIVARIATE displays extreme observations, missing values, and other statistics for the variables included in the VAR statement.

```
proc univariate data=orion.nonsales2;
    var Salary;
run;
```

PROC UNIVARIATE DATA=*SAS-data-set*;
    <VAR *variable(s)*;>
RUN;

If the VAR statement is omitted, PROC UNIVARIATE analyzes all numeric variables in the data set.

Viewing the Output: Extreme Observations

The *Extreme Observations* section includes the five lowest and five highest values for the analysis variable and the corresponding observation numbers.   Partial PROC UNIVARIATE Output

```
                    Extreme Observations

    -----Lowest----              -----Highest----

    Value        Obs              Value        Obs

    2401          20             163040          1
    2650          13             194885        231
   24025          25             207885         28
   24100          19             268455         29
   24390         228             433800         27
```

Not conf hot our observation #

✎ Obs is the observation number, not the count of observations with that value.

§.sas

NEXTROBS= Option

The *NEXTROBS= option* specifies the number of extreme observations to be displayed.

```
proc univariate data=orion.nonsales2
                 nextrobs=3;
   var Salary;
run;
```

Partial PROC UNIVARIATE Output

```
              The UNIVARIATE Procedure
                Variable:  Salary

                Extreme Observations

      -----Lowest----          -----Highest----

      Value      Obs           Value        Obs

       2401       20          207885         28
       2650       13          268455         29
      24025       25          433800         27
```

§sas

ID Statement

The *ID statement* displays the value of the identifying variable (or variables) in addition to the observation number.

```
proc univariate data=orion.nonsales2;
   var Salary;
   id Employee_ID;
run;
```

ID *variable(s)*;

Viewing the Output

Partial PROC UNIVARIATE Output

```
                    The UNIVARIATE Procedure
                       Variable:  Salary

                     Extreme Observations

-----------Lowest----------          -----------Highest----------

Value     Employee_ID    Obs          Value     Employee_ID    Obs

 2401        120191       20         163040        120101        1
 2650        120115       13         194885        121141      231
24025        120196       25         207885        120260       28
24100        120190       19         268455        120262       29
24390        121132      228         433800        120259       27
```

§sas

Viewing the Output: Missing Values Section

The *Missing Values* section displays the number and percentage of observations with missing values for the analysis variable.

Partial PROC UNIVARIATE Output

```
                        Missing Values

                                    -----Percent Of-----
        Missing                                 Missing
        Value       Count       All Obs             Obs

            .           1          0.43          100.00
```

§.sas

Short Answer Poll

PROC UNIVARIATE identified two observations
with **Salary** values less than 24,000.

What procedure can be used to display the observations that contain the
invalid values?

Short Answer Poll – Correct Answer

PROC UNIVARIATE identified two observations
with **Salary** values less than 24,000.

What procedure can be used to display the observations that contain the
invalid values? **PROC PRINT**

```
proc print data=orion.nonsales2;
    where Salary<24000;
run;
```

```
Obs    Employee_ID    First      Last          Gender    Salary    Job_Title             Country

  4       120106       John       Hornsey          M         .      Office Assistant II      AU
 13       120115       Hugh       Nichollas        M       2650     Service Assistant I      AU
 20       120191       Jannene    Graham-Rowe      F       2401     Trainee                  AU
```

Analyzing data with PROC UNIVARIATE

This demonstration shows the full output produced by the UNIVARIATE procedure.

21-univariate

Producing Descriptive Statistics

The TABULATE Procedure

The TABULATE Procedure

The TABULATE procedure displays descriptive statistics in tabular format.

General form of the TABULATE procedure:

```
PROC TABULATE DATA=SASdataset <options>;
     CLASS classificationvariable(s);
     VAR analysis-variable(s);
     TABLE        pageexpression,
                  rowexpression,
                  columnexpression </ option(s)>;
RUN;
```

None of these are optional (handwritten annotation)

Dimensional Tables

The TABULATE procedure produces one-, two-, or three-dimensional tables.

	page dimension	row dimension	column dimension
one-dimensional			✔
two-dimensional		✔	✔
three-dimensional	✔	✔	✔

One-Dimensional Table

Country	
AU	US
N	N
63	102

- **`Country`** is in the column dimension.

Two-Dimensional Table

	Country	
	AU	US
	N	N
Gender		
F	27	41
M	36	61

- **Country** is in the column dimension.

- **Gender** is in the row dimension.

Three-Dimensional Table

Job_Title Sales Rep. I		
	Country	
	AU	US
	N	N
Gender		
F	8	13
M	13	29

- **Country** is in the column dimension.

- **Gender** is in the row dimension.

- **Job_Title** is in the page dimension.

The TABLE Statement

The TABLE statement describes the structure of the table.

| table | **page** , expression | **row** , expression | **column** ; expression |

dimension expressions
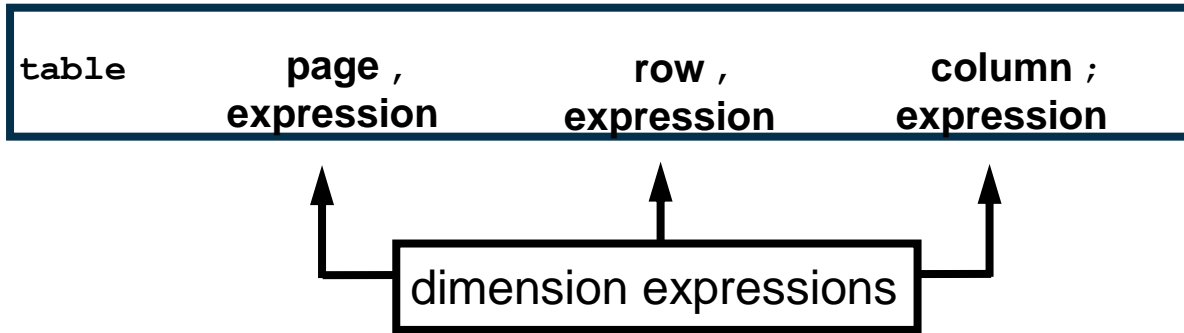
- Commas separate the dimension expressions.
- <u>Every</u> variable that is part of a dimension expression must be specified as a classification variable (CLASS statement) or an analysis variable (VAR statement).

The TABLE Statement

PROC   TABULATE

age   w   olumn

The TABLE Statement

table	**page** , **expression**	**row** , **expression**	**column** ; **expression**

Examples:

```
table Country;
```

```
table Gender , Country;
```

```
table Job_Title , Gender , Country;
```

§.sas

The CLASS Statement

The CLASS statement identifies variables to be used as classification, or grouping, variables.

General form of the CLASS statement:

CLASS *classificationvariable(s)*;

- N, the number of nonmissing values, is the default statistic for classification variables.

- Examples of classification variables:

  `Job_Title`, `Gender`, and `Country`

The VAR Statement

The VAR statement identifies the numeric variables for which statistics are calculated.

General form of the VAR statement:

VAR *analysisvariable(s)*;

- SUM is the default statistic for analysis variables.
- Examples of analysis variables:

    Salary and **Bonus**

One-Dimensional Table

```
proc tabulate data=orion.sales;
    class Country;
    table Country;
run;
```

Country	
AU	US
N	N
63	102

§sas

Two-Dimensional Table

```
proc tabulate data=orion.sales;
    class Gender Country;
    table Gender, Country;
run;
```

	Country	
	AU	US
	N	N
Gender		
F	27	41
M	36	61

Three-Dimensional Table

```
proc tabulate data=orion.sales;
    class Job_Title Gender Country;
    table Job_Title, Gender, Country;
run;
```

Job_Title Sales Rep. I

	Country	
	AU	US
	N	N
Gender		
F	8	13
M	13	29

Job_Title Sales Rep. II

	Country	
	AU	US
	N	N
Gender		
F	10	14
M	8	14

Job_Title Sales Rep. III

	Country	
	AU	US
	N	N
Gender		
F	7	8
M	10	9

Dimension Expression

Elements that can be used in a dimension expression:

- classification variables

- analysis variables

- the universal class variable ALL

- keywords for statistics

Operators that can be used in a dimension expression:

- blank, which concatenates table information

- asterisk *, which crosses table information

- parentheses (), which group elements

Dimension Expression

```
proc tabulate data=orion.sales;
    class Gender Country;
    var Salary;
    table Gender all, Country*Salary;
run;
```

	Country	
	AU	US
	Salary	Salary
	Sum	Sum
Gender		
F	747965.00	1207900.00
M	1152050.00	2033505.00
All	1900015.00	3241405.00

§.sas

PROC TABULATE Statistics

Descriptive Statistic Keywords

	CSS	CV	LCLM	MAX
MEAN	MIN	MODE	N	NMISS
KURTOSIS	RANGE	SKEWNESS	STDDEV	STDERR
SUM	SUMWGT	UCLM	USS	VAR
PCTN	REPPCTN	PAGEPCTN	ROWPCTN	COLPCTN
PCTSUM	REPPCTSUM	PAGEPCTSUM	ROWPCTSUM	COLPCTSUM

Quantile Statistic Keywords

MEDIAN | P50	P1	P5	P10	Q1 | P25
Q3 | P75	P90	P95	P99	QRANGE

Hypothesis Testing Keywords

PROBT	T			

PROC TABULATE Statistics

Format
*f= Format

```
proc tabulate data=orion.sales;
    class Gender Country;
    var Salary;
    table Gender all, Country*Salary*(min max)*f=dollar12.;
run;
```

	Country			
	AU		US	
	Salary		Salary	
	Min	Max	Min	Max
Gender				
F	$25,185	$30,890	$25,390	$83,505
M	$25,745	$108,255	$22,710	$243,190
All	$25,185	$108,255	$22,710	$243,190

Additional SAS Statements

Additional statements can be added to enhance the report.

```
proc format;
    value $ctryfmt 'AU'='Australia'
                   'US'='United States';
run;

options nodate pageno=1;

proc tabulate data=orion.sales;
    class Gender Country;
    var Salary;
    table Gender all, Country*Salary*(min max);
    where Job_Title contains 'Rep';
    label Salary='Annual Salary';
    format Country $ctryfmt.;
    title 'Sales Rep Tabular Report';
run;
```

Additional SAS Statements

HTML Output

Sales Rep Tabular Report

	Country			
	Australia		United States	
	Annual Salary		Annual Salary	
	Min	Max	Min	Max
Gender				
F	25185.00	30890.00	25390.00	32985.00
M	25745.00	36605.00	22710.00	35990.00
All	25185.00	36605.00	22710.00	35990.00

Output Data Sets

PROC TABULATE produces output data sets using the following method:

PROC TABULATE DATA=_SAS-data-set_
                          OUT=_SAS-data-set_ _<options>_**;**

The output data set contains the following variables:

- BY variables

- class variables

- the automatic variables **_TYPE_**, **_PAGE_**, and **_TABLE_**

- calculated statistics

PROC Statement OUT= Option

```
proc tabulate data=orion.sales
               out=work.tabulate;
   where Job_Title contains 'Rep';
   class Job_Title Gender Country;
   table Country;
   table Gender, Country;
   table Job_Title, Gender, Country;
run;

proc print data=work.tabulate;
run;
```

PROC Statement OUT= Option
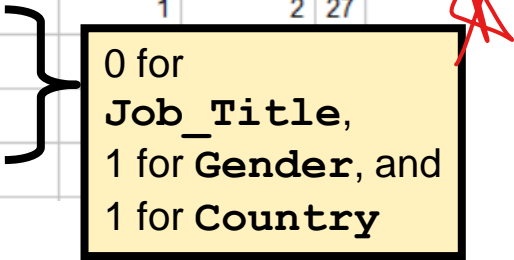
Partial PROC PRINT Output

Obs	Job_Title	Gender	Country	_TYPE_	_PAGE_	_TABLE_	N
1			AU	001	1	1	61
2			US	001	1	1	98
3		F	AU	011	1	2	27
4		F	US	011	1	2	40
5		M	AU	011	1	2	34
6		M	US	011	1	2	58
7	Sales Rep. I	F	AU	111	1	3	8
8	Sales Rep. I	F	US	111	1	3	13
9	Sales Rep. I	M	AU	111	1	3	13
10	Sales Rep. I	M	US	111	1	3	29
11	Sales Rep. II	F	AU	111	2	3	10
12	Sales Rep. II	F	US	111	2	3	14
13	Sales Rep. II	M	AU	111	2	3	8
14	Sales Rep. II	M	US	111	2	3	14

PROC Statement OUT= Option

TYPE is a character variable that shows which combination of class variables produced the summary statistics in that observation.

Partial PROC PRINT Output

Obs	Job_Title	Gender	Country	_TYPE_	_PAGE_	_TABLE_	N
1			AU	001	1	1	61
2			US	001	1	1	98
3		F	AU	011	1	2	27
4		F	US	011			
5		M	AU	011			
6		M	US	011			

0 for **Job_Title**, 1 for **Gender**, and 1 for **Country**

PROC Statement OUT= Option

PAGE is a numeric variable that shows the logical page number that contains that observation.

Partial PROC PRINT Output

7	Sales Rep. I	F	AU	111	1		3	8
8	Sales Rep. I	F	US	111	1			
9	Sales Rep. I	M	AU	111	1			
10	Sales Rep. I	M	US	111	1		3	29
11	Sales Rep. II	F	AU	111	2		3	10
12	Sales Rep. II	F	US	111	2			
13	Sales Rep. II	M	AU	111	2			
14	Sales Rep. II	M	US	111	2		3	14
15	Sales Rep. III	F	AU	111	3		3	7
16	Sales Rep. III	F	US	111	3			
17	Sales Rep. III	M	AU	111	3			
18	Sales Rep. III	M	US	111	3		3	9

Page 1 for Sales Rep. I

Page 2 for Sales Rep. II

Page 3 for Sales Rep. III

§sas

PROC Statement OUT= Option

`_TABLE_` is a numeric variable that shows the number of the TABLE statement that contains that observation.

Partial PROC PRINT Output

Obs	Job_Title	Gender	Country	_TYPE_	_PAGE_	_TABLE_	N
1						1	61
2						1	98
3		F	AU	011	1	2	27
4						2	40
5		M		011		2	34
6		M	US	011	1	2	58
7	Sales Rep. I	F	AU	111	1	3	8
8	Sales Rep.					3	13
9	Sales Rep.	M		111		3	13
10	Sales Rep. I	M	US	111	1	3	29

1 for first TABLE statement

2 for second TABLE statement

3 for third TABLE statement