22S:166
SAS

Queries and SQL
More on Data Integrity

Lecture 20
Nov. 26, 2007

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

Structured Query Language

- query: a view of data which represents the data from one or more tables
- queries built in a relational database using Structured Query Language or SQL
- SQL is the standard language for relational databases
- includes the capability of manipulating both the structure of a database and its data
- most common use: to create a simple SELECT query

Proc sql in SAS

- SAS data files and SQL tables
  - structure of an SQL table is very similar to that of a SAS data file
  - only difference: SASdata file has inherent ordering
  - in SAS System, SQL table is represented as a SASdata file
- *proc sql* can perform some of the operations provided y the *data* step and the *print, sort,* and *means* procedures
  - often can achieve same results as these procedures with fewer and short statements
  - why should you still know how to do these tasks with *print, sort, means*, etc.?
    * because you are likely to have to maintain or modify older programs written before *proc sql* was added to SAS

Queries using proc sql select statement

- *select* statement in *proc sql* finds and displays specified records and variables
- can also link files, calculate summary statistics, sort, etc.

Return to sites and deposition example

```
options linesize = 75 pagesize = 60 nodate nonumber ;

data depo ;
infile 'depoRep90s.asp' firstobs = 8 ;
input SiteID  $ Per $8. Year  Crit1 Crit2 Crit3 Crit4 Ca Mg
K  Na  NH4  NO3 InorgN Cl SO4 HLab  HField  Svol  Ppt  Pct  ValidF ValidL
Days @196 Date1 mmddyy10. @209 Date2 mmddyy10.  ;
drop Per Crit1-Crit4  Ca Mg K  Na  NH4  NO3 InorgN Cl HLab  HField
Svol  Ppt  Pct  ValidF ValidL ;
daysop = Date2 - Date1 ;
format Date2 Date1 date8. ;
run ;

data sites ;
infile '/space/kcowles/166/lectures/lect1mkc/stateCO.asp' firstobs = 19
missover ;
input @13 SiteID $ @20 sitename $18. @40 strtdate mmddyy10. @53 stopdate
if strtdate ne . ;   * subsetting if: exclude observations meeting condit
format strtdate stopdate date8. ;
drop sitename ;
run ;
```


- sites file
  - SiteID
  - strtdate
  - stopdate
  - elev
  - SiteID
- depo file
  - SiteID
  - Year
  - SO4


```
proc sql ;
    title 'Proc sql listings' ;
    select * from sites ;            /* list all variables and records *
```

Proc sql listings

SiteID	strtdate	stopdate	elev
C000	22APR80	.	2298
C001	04OCT83	.	1213
C002	05JUN84	.	3520
C008	29DEC87	.	2502
C010	02FEB99	.	2926
C015	20MAR79	.	1998
C019	29MAY80	.	2490
C021	17OCT78	.	2362
C022	22MAY79	.	1641
C091	26MAY92	.	3292
C092	13JAN88	.	3206
C093	14OCT86	.	2527
C094	04NOV86	.	2524
C095	29JUL86	02JAN90	2758
C096	29JUL86	.	3249
C097	07FEB84	.	3234
C098	16AUG83	.	3159
C099	28APR81	.	2172


```
    select SiteID, elev from sites ;  /* list selected variables, all rec
```

Proc sql listings

SiteID	elev
C000	2298
C001	1213
C002	3520
C008	2502
C010	2926
C015	1998
C019	2490
C021	2362
C022	1641
C091	3292
C092	3206
C093	2527
C094	2524
C095	2758
C096	3249
C097	3234
C098	3159
C099	2172

```
* multiple-table query ;

title2 'Multiple table query' ;
select s.siteID, s.elev, d.SO4, d.Year
from sites s, depo d
where s.SiteID = d.SiteID
order by s.SiteID ;
```

Multiple table query

SiteID	elev	SO4	Year
CD00	2298	1.2	1992
CD00	2298	1.28	1998
CD00	2298	1.07	1996
CD00	2298	2.08	1991
CD00	2298	1.01	1999
CD00	2298	1.18	2000
CD00	2298	1.1	1997
CD00	2298	1.5	1993
CD00	2298	1.46	1995
CD00	2298	1.31	1994
CD01	1213	3.64	1995
CD01	1213	3.09	1992
CD01	1213	2.98	1994
CD01	1213	2.3	1993
CD01	1213	2.44	1998
CD01	1213	2.53	1997
CD01	1213	3.19	1991
CD01	1213	2.06	2000
.			
.			
.			


```
* more sophisticated query ;

title2 'More complicated SELECT and ORDER' ;
select s.siteID, s.elev, d.SO4, d.Year
from sites s, depo d
where s.SiteID = d.SiteID and d.Year > 1995
order by s.SiteID,  d.Year ;
```

More complicated SELECT and ORDER

SiteID	elev	SO4	Year
CD00	2298	1.07	1996
CD00	2298	1.1	1997
CD00	2298	1.28	1998
CD00	2298	1.01	1999
CD00	2298	1.18	2000
CD01	1213	2.99	1996
CD01	1213	2.53	1997
CD01	1213	2.44	1998
CD01	1213	3.49	1999
CD01	1213	2.06	2000
CD02	3520	14.82	1996
CD02	3520	10.9	1997
CD02	3520	8.67	1998
CD02	3520	10.7	1999
CD02	3520	19.32	2000
.			
.			
.			


```
* summing and grouping ;

title2 'Total Deposition' ;

select siteID, sum(SO4) as totso4
from depo
group by SiteID
order by SiteID
;
```

Proc sql listings
Total Deposition

SiteID	totso4
CD00	13.19
CD01	28.71
CD02	115.68
CD08	26.25
CD10	4.85
CD15	23.28
CD19	26.17
CD21	35.49
CD22	30.16
CD91	69.82
CD92	33.01
CD93	65.73
CD94	40.81
CD96	39.53
CD97	82.16
CD98	57.27
CD99	39.88

Producing report with proc means

```
proc means data = depo ;
class SiteID ;           /* separate summary stats by this variable */
var SO4 ;                /* which numeric variable to summarize */
output out = meandepo mean=avgso4 ;  /* identify output dataset and
                                        variable name for summary stat */
run ;

proc print data = meandepo ;
run ;
```

Obs	Site ID	_TYPE_	_FREQ_	avgso4
1		0	161	4.5465
2	CD00	1	10	1.3190
3	CD01	1	10	2.8710
4	CD02	1	10	11.5680
5	CD08	1	10	2.6250
6	CD10	1	2	2.4250
7	CD15	1	10	2.3280
8	CD19	1	10	2.6170
9	CD21	1	10	3.5490
10	CD22	1	10	3.0160
11	CD91	1	9	7.7578
12	CD92	1	10	3.3010
13	CD93	1	10	6.5730
14	CD94	1	10	4.0810
15	CD96	1	10	3.9530
16	CD97	1	10	8.2160
17	CD98	1	10	5.7270
18	CD99	1	10	3.9880