NCHS Research Data Center

CDC/NCHS
Research Data Center

Presented July 23, 2003
Bureau of Transportation Statistics
Confidentiality Seminar Series

Kenneth W. Harris
Acting Director
(301) 458-4262
Kwh1@cdc.gov

Vijay Gambhir
Computer Scientist
(301) 458-4226
Vgambhir@cdc.gov

Operational Comparison of the NCHS Research Data Center and the Census Research Data Center

	NCHS RDC	Census RDC
Data available	Virtually any NCHS survey without direct identifiers	Title 13 data such as CPS, MEPS, and others
Researcher-supplied data	Allowed	Allowed
Research proposals	Required	Required
Review cycle	Continuous	Three times per year
Turn-around time	2-3 weeks	4-8 months
Tenure at RDC	Short term or long term (minimum charge of 2 days)	Minimum 3 months
Remote Access	Yes	No

	NCHS RDC	Census RDC
Type of projects	Tabular or model-based	Model-based only
Researcher costs		
Remote access	$500/month if file size < 130,000 records. $1000/month if file size > 130,000 records.	N/A
On site	$200/day	

Operational Comparison of the NCHS Research Data Center and the Census Research Data Center

File set up	$500/day's effort	N/A
Computer equipment		
Hardware	Windows NT server/Windows 2000 workstations	Unix server As described in Reznek talk
Software	SAS/Fortran/Stata Others available upon request	SAS/Stata Others available upon request

Note: At this time files containing restricted or confidential data cannot be transmitted across data center boundaries.

Analytic Data Research by Email (ANDRE)

- NCHS has been providing remote data access to the researchers through ANDRE since April 1998
- In the past five years ANDRE has served 45 different data analysts and executed over 10,000 SAS programs for their research programs.

Main Features of ANDRE

- Completely automated system. Operates round the clock without any human intervention
- Registered subscribers only.
  1. Proposals already reviewed and approved
  2. Have an agreement with NCHS/RDC
- Unlimited access during the subscription period

Data Requests

- Registered user can submit data requests by email from anywhere and at any time.
- Results of the data request released to a specified email address that has been certified to be secure by the subscriber and approved by NCHS/RDC

Authentication

- Multi-levels of system security:
  - Submission syntax
  - User id
  - Password
  - Email/code word
  - Package
  - Path info

Data Request Analysis

- Compliance with the disclosure limitation constraints of NCHS
- Integrity of the system
  - Resource constraints (CPU time & Storage requirements)
  - Protection of ANDRE's work environment

Prevention of Direct Disclosure

- Cleaning up of the Log File
- Categorization of SAS commands/words
  1. Forbidden Commands
  2. Modifications to the Commands
  3. Output suppression

Sample:- Original Log

```
1    options nocenter;
2    Data one;
3    Infile 'd:\nchs\respnd95.dat' lrecl=13064;
4    Input
5    TODAYSPG 6847-6847
6    CONSTAT1 11934-11935
7    CONSTAT2 11936-11937
8    CONSTAT3 11938-11939
9    CONSTAT4 11940-11941
10    SEX1MTHD 11945-11946
11    POST_WT  12350-12359;
12    if constat1 = 'ab' then vjvar=1; else vjvar = 2;
13    WGT1000=POST_WT/1000;
14    title 'NSFG cycle 1995';

NOTE: Character values have been converted to numeric values at the places given by: (Line):(Column).
      12:15
NOTE: The infile 'd:\nchs\respnd95.dat' is:
      File Name=d:\nchs\respnd95.dat,
      RECFM=V,LRECL=13064

NOTE: Invalid numeric data, 'ab' , at line 12 column 15.
RULE:    ----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+----8----+----9---
     -+----0
1      1000000111260837511521 1                    1050  12     1069211241124111 89
   101                                          2
   201     19211059110611197
                          .....
                          .....
                          .....
```

Sample:- Original Log (cont.)

.....
.....

12901 11232521101 05267213103033921811931011103
0103000000032112000039270221061151120040 3 1344 1316
13001 622501001006034
TODAYSPG=1 CONSTAT1=5 CONSTAT2=88 CONSTAT3=88 CONSTAT4=88 SEX1MTHD=1
POST_WT=2545.7569 vjvar=2 WGT1000=2.5457569 _ERROR_=1
_N_=20

NOTE: 10847 records were read from the infile 'd:\nchs\respnd95.dat'.
The minimum record length was 13064.
The maximum record length was 13064.
NOTE: The data set WORK.ONE has 10847 observations and 9 variables.
NOTE: DATA statement used:
real time 39.88 seconds
cpu time 12.10 seconds

15 proc freq;
16 tables CONSTAT1 vjvar;
17 run;

NOTE: There were 10847 observations read from the data set WORK.ONE.
NOTE: PROCEDURE FREQ used:
real time 0.49 seconds
cpu time 0.04 seconds

Sample:- Cleaned Log

```
1    options nocenter;
2    Data one;
3    Infile 'd:\nchs\respnd95.dat' lrecl=13064;
4    Input
5    TODAYSPG 6847-6847
6    CONSTAT1 11934-11935
7    CONSTAT2 11936-11937
8    CONSTAT3 11938-11939
9    CONSTAT4 11940-11941
10   SEX1MTHD 11945-11946
11   POST_WT  12350-12359;
12   if constat1 = 'ab' then vjvar=1; else vjvar = 2;
13   WGT1000=POST_WT/1000;
14   title 'NSFG cycle 1995';

NOTE: Character values have been converted to numeric values at the places given by: (Line):(Column).
      12:15
NOTE: The infile 'd:\nchs\respnd95.dat' is:
      File Name=d:\nchs\respnd95.dat,
      RECFM=V,LRECL=13064
```

NOTE: Invalid numeric data, 'ab' , at line 12 column 15.

Sample:- Cleaned Log (cont.)

NOTE: 10847 records were read from the infile 'd:\nchs\respnd95.dat'.
The minimum record length was 13064.
The maximum record length was 13064.
NOTE: The data set WORK.ONE has 10847 observations and 9 variables.
NOTE: DATA statement used:
real time 39.88 seconds
cpu time 12.10 seconds

```
15   proc freq;
16   tables CONSTAT1 vjvar;
17   run;
```

NOTE: There were 10847 observations read from the data set WORK.ONE.
NOTE: PROCEDURE FREQ used:
real time 0.49 seconds
cpu time 0.04 seconds

Forbidden Commands

- Commands that pose unacceptable disclosure risks OR
- Disallowed to protect integrity/internal environment of ANDRE

Add	editor	report	iml
Print	first.	Pctn	nofreq
Obs	last.	Pctsum	nocum
Firstobs	nocol	tabulate	editor
Browse	summary	list	put

Commands Modification

- Modify user's program to enforce restrictions on options allowed with certain SAS procedures to prevent objectionable info appearing in the output

```
PROC MEANS  n mean std;
```

Output Suppression

- Wiping out of extreme values from the output of Proc Univariate
- Suppressing complete output line (Procs Means, corr, Univariate etc) where sample size less than the minimum acceptable value.

Proc Means Suppression

The MEANS Procedure

Variable	Label	N	Mean	Std Dev
EXPEND_R	Current expend/pupil in public schl/1000	5424	5.0830820	1.3958710
	***** Values Suppressed *****			
RPUB87	exp. for contr. serv. and supplies 1997$	5424	23472052.60	18806802.86
RPUB92	exp. for contr. serv. and supplies 1997$	5424	34800922.98	30481634.59
PRGPRO	Coordinated Pregnancy Prevention Program	1708	0.0679157	0.2516749
HIVED	HIV/AIDS Education	1708	3.5146370	0.8044378
	***** Values Suppressed *****			
PRGPRO87	Coordinated Pregnancy Prevention Program	5424	0.0540192	0.2260764
HIVED87	HIV/AIDS Education	5424	3.4968658	0.8008324
WT_PER15	% Wt females aged 15-19/total 15-19	5424	0.7279681	0.1265796
BK_PER15	% Bk females aged 15-19/total 15-19	5424	0.1409869	0.0932332
HS_PER15	% Hs females aged 15-19/total 15-19	5424	0.0962413	0.1055191
TEENMMC2	Teenmom by cohort (1,2,3r)	1201	1.7119067	0.7715351
C18_2_1S	R in C2 (vs 1) at 18-19 endpt (1,2)	1770	1.5248588	0.4995228
TM2_1S18	R tnmm in Coh 2 (vs 1)-age 18 @ ext	358	1.4804469	0.5003168
AGE_12	Date R = 12 in century months	6450	979.5613953	69.3124265
STRTST	IA5 Date R started living in current sta	3870	1132.55	753.2066507
BDAYCENM	R date of birth	6450	835.5613953	69.3124265
RAVPAY95	real av. an. pay 95 dollars	5424	26933.93	2826.80
PERCAFDC	percent of households receiving AFDC	5424	0.0422254	0.0127307
SALARY	teacher salaries real 96-97$$$	5424	35338.66	5729.11

Proc Univariate Output Unsuppressed

The SAS System 9
14:09 Sunday, October 24, 1999

Univariate Procedure

Variable=AVHRATET

Moments

N	2283	Sum Wgts	2283
Mean	-4.66219	Sum	-10643.8
Std Dev	1.892017	Variance	3.57973
Skewness	-2.11919	Kurtosis	6.892929
USS	57792.36	CSS	8168.944
CV	-40.5821	Std Mean	0.039598
T:Mean=0	-117.738	Pr>|T|	0.0001
Num ^= 0	2283	Num > 0	0
M(Sign)	-1141.5	Pr>=|M|	0.0001
Sgn Rank	-1303593	Pr>=|S|	0.0001

Quantiles(Def=5)

100% Max	-0.25314	99%	-1.62008
75% Q3	-3.56179	95%	-2.37588
50% Med	-4.50491	90%	-2.79152
25% Q1	-5.30374	10%	-6.07639
0% Min	-13.5463	5%	-7.19645
		1%	-12.7402
Range	13.29321		
Q3-Q1	1.741949		
Mode	-13.5463		

Extremes

Lowest	Obs	Highest	Obs
-13.5463	(1547)	-0.90519	(649)
-13.5397	(1836)	-0.81756	(1094)
-13.4637	(2084)	-0.76928	(1739)
-13.4413	(1127)	-0.5907	(21)
-13.4402	(1088)	-0.25314	(400)

Proc Univariate Output Suppressed

The SAS System 9

14:09 Sunday, October 24, 1999

Univariate Procedure

Variable=AVHRATET

Moments

N	2283	Sum Wgts	2283
Mean	-4.66219	Sum	-10643.8
Std Dev	1.892017	Variance	3.57973
Skewness	-2.11919	Kurtosis	6.892929
USS	57792.36	CSS	8168.944
CV	-40.5821	Std Mean	0.039598
T:Mean=0	-117.738	Pr>|T|	0.0001
Num ^= 0	2283	Num > 0	0
M(Sign)	-1141.5	Pr>=|M|	0.0001
Sgn Rank	-1303593	Pr>=|S|	0.0001

Quantiles(Def=5)

100% Max	-0.25314	99%	-1.62008
75% Q3	-3.56179	95%	-2.37588
50% Med	-4.50491	90%	-2.79152
25% Q1	-5.30374	10%	-6.07639
0% Min	-13.5463	5%	-7.19645
		1%	-12.7402
Range	13.29321		
Q3-Q1	1.741949		
Mode	-13.5463		

Proc Univariate Output Suppressed (sample size = 1)

Univariate Procedure

Variable=FREQ (sum) freq

Moments **Quantiles(Def=5)**

Serious Disclosure limitation Violations

Values too low to release

Output of Proc Univariate withheld

Proc Freq Suppression (one way Tables)

- Suppress at least two consecutive rows to prevent derivation of suppressed values from cumulative totals.
- Disallow single row output.

1-Way Freq Table suppressed

LOGRNTOPAT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0.2277839309	**?????**	**?????**	**?????**	**?????**
0.2277839309	**?????**	**?????**	**?????**	**?????**
0.2305236586	5	0.08	6429	97.99
0.231111721	5	0.08	6434	98.06
0.232058915	**?????**	**?????**	**?????**	**?????**
0.232058915	**?????**	**?????**	**?????**	**?????**
0.2436220827	**?????**	**?????**	**?????**	**?????**
0.2436220827	**?????**	**?????**	**?????**	**?????**
0.2498117984	6	0.09	6456	98.40
0.2504106777	6	0.09	6462	98.49
0.2513144283	18	0.27	6480	98.77
0.2595111955	6	0.09	6486	98.86
0.2670627852	**?????**	**?????**	**?????**	**?????**
0.2670627852	**?????**	**?????**	**?????**	**?????**
0.2736958305	5	0.08	6500	99.07
0.2814124594	5	0.08	6505	99.15
0.3022808719	6	0.09	6511	99.24
0.3364722366	10	0.15	6521	99.39

1-Way Freq Table suppressed (cont.)

LOGRNTOPAT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0.3403258059	**?????**	**?????**	**?????**	**?????**
0.3403258059	**?????**	**?????**	**?????**	**?????**
0.3715635564	6	0.09	6537	99.63
0.3856624808	**?????**	**?????**	**?????**	**?????**
0.3856624808	**?????**	**?????**	**?????**	**?????**
0.6931471806	6	0.09	6550	99.83
1.2527629685	**?????**	**?????**	**?????**	**?????**
1.2527629685	**?????**	**?????**	**?????**	**?????**
1.2527629685	**?????**	**?????**	**?????**	**?????**

Proc Freq Suppression (Two way Tables)

- Rows and columns totals preserved
- Cells with values less than the acceptable minimum are suppressed
- Additional suppressions to ensure that no row and no column has single suppression.
- Logical stitching of horizontal and vertical splits.

Proc Freq: 2-way Tables Suppression

TABLE OF FAMREL BY FAMSIZER

FAMREL / Frequency / Percent / Row Pct / Col Pct	2	3	4	5	Total
3	94	388	792	533	2206
	3.97	16.40	33.47	22.53	93.24
	4.26	17.59	35.90	24.16	
	98.95	96.28	96.12	94.34	
4	??????	9	22	27	104
	??????	0.38	0.93	1.14	4.40
	??????	8.65	21.15	25.96	
	??????	2.23	2.67	4.78	
6	??????	6	10	5	56
	??????	0.25	0.42	0.21	2.37
	??????	10.71	17.86	8.93	
	??????	1.49	1.21	0.88	
Total	95	403	824	565	2366
	4.02	17.03	34.83	23.88	100.00

Proc Freq: 2-way Tables Suppression (Cont.)

checking frequencies 4

12:01 Thursday, May 6, 1999

TABLE OF FAMREL BY FAMSIZER

FAMREL / Frequency / Percent / Row Pct / Col Pct	FAMSIZER 6	7	8	9	Total
3	209	98	19	73	2206
	8.83	4.14	0.80	3.09	93.24
	9.47	4.44	0.86	3.31	
	90.48	83.05	59.38	74.49	
4	13	10	??????	12	104
	0.55	0.42	??????	0.51	4.40
	12.50	9.62	??????	11.54	
	5.63	8.47	??????	12.24	
6	9	10	??????	13	56
	0.38	0.42	??????	0.55	2.37
	16.07	17.86	??????	23.21	
	3.90	8.47	??????	13.27	
Total	231	118	32	98	2366

Contact Information

For general Questions/Comments

Email : rdca@cdc.gov Phone: (301) 458-4732

For On-site Info:

Email : Neb9@cdc.gov Phone: (301) 458-4097

For Remote Access Info:

Email : vgambhir@cdc.gov Phone: (301) 458-4226