

**22S:166**  
**Computing in Statistics**

**Data validation and description**  
**Proc format**

Lecture 18  
Nov. 3, 2006

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

**Data checking and screening**

- important to do prior to any multivariate analyses
- purpose: to identify incorrect, invalid, or otherwise suspect data
- begin with simple descriptive statistics and plots for each variable
- types of checks for binary, nominal, and ordinal data
  - frequency and proportion of invalid categories
  - frequency and proportion of missing classifications
  - adequate representation of categories of interest?

- types of checks for continuous data
  - range screens
  - consistency screens
  - accuracy of measurement

Primary question for the applied statistician:  
Does this make sense?

**Example: the Berkeley Guidance Study**

```
options linesize = 70 pagesize = 60 nodate nonumber ;

data berkboy ;
infile '/group/ftp/pub/kcowles/datasets/berkboy.dat' ;
input wt2 ht2 wt9 ht9 lg9 st9 wt18 ht18 lg18 st18 soma ;
run ;

proc corr ;
run ;

proc reg data = berkboy ;
model soma = ht2 wt2 ht9 wt9 st9 ;
run ;

proc reg data = berkboy ;
model soma = ht9 wt9 st9 ;
run ;
```

The CORR Procedure

11 Variables: wt2 ht2 wt9 ht9 lg9 st9  
wt18 ht18 lg18 st18 soma

st18 31.00000 44.10000  
soma 152.00000 252.00000

Simple Statistics

Variable	N	Mean	Std Dev	Sum
wt2	26	214.53846	8.37726	5578
ht2	26	13.59231	1.61862	353.40000
wt9	26	88.40000	3.03592	2298
ht9	26	31.58462	4.35850	821.20000
lg9	26	136.54615	5.31603	3550
st9	26	27.53077	1.89626	715.80000
wt18	26	71.30769	10.69119	1854
ht18	26	71.58077	11.56509	1861
lg18	26	180.03846	6.39619	4681
st18	26	36.33846	2.72882	944.80000
soma	26	210.42308	25.26210	5471

Simple Statistics

Variable	Minimum	Maximum
wt2	201.00000	228.00000
ht2	11.30000	17.20000
wt9	81.30000	92.20000
ht9	24.50000	43.10000
lg9	125.40000	146.00000
st9	24.20000	32.40000
wt18	45.00000	98.00000
ht18	50.30000	110.20000
lg18	169.40000	195.10000

Pearson Correlation Coefficients, N = 26  
Prob > |r| under H0: Rho=0

	wt2	ht2	wt9	ht9	lg9	st9
wt2	1.00000	0.09354 0.6495	0.28546 0.1575	-0.07875 0.7022	0.08547 0.6781	-0.07209 0.7264
ht2	0.09354 0.6495	1.00000	0.49768 0.0097	0.57922 0.0019	0.38230 0.0539	0.58066 0.0019
wt9	0.28546 0.1575	0.49768 0.0097	1.00000	0.53122 0.0052	0.77583 <.0001	0.28446 0.1590
ht9	-0.07875 0.7022	0.57922 0.0019	0.53122 0.0052	1.00000	0.62049 0.0007	0.90553 <.0001
lg9	0.08547 0.6781	0.38230 0.0539	0.77583 <.0001	0.62049 0.0007	1.00000	0.35332 0.0766

The REG Procedure  
Model: MODEL1  
Dependent Variable: soma

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2295.78296	459.15659	0.67	0.6491
Error	20	13659	682.92816		
Corrected Total	25	15954			

  

Root MSE	26.13289	R-Square	0.1439
Dependent Mean	210.42308	Adj R-Sq	-0.0701
Coeff Var	12.41921		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	51.66778	293.72149	0.18	0.8621
ht2	1	0.42206	4.47469	0.09	0.9258
wt2	1	-0.22891	0.70601	-0.32	0.7491
ht9	1	0.29498	4.16397	0.07	0.9442
wt9	1	1.20330	3.00337	0.40	0.6929
st9	1	3.13973	8.73447	0.36	0.7230

The REG Procedure  
Model: MODEL1  
Dependent Variable: soma

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2216.26058	738.75353	1.18	0.3390
Error	22	13738	624.45844		
Corrected Total	25	15954			

Root MSE	24.98917	R-Square	0.1389
Dependent Mean	210.42308	Adj R-Sq	0.0215
Coeff Var	11.87568		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	33.23337	259.54880	0.13	0.8993
ht9	1	0.68933	3.65283	0.19	0.8520
wt9	1	0.90587	2.32088	0.39	0.7001
st9	1	2.73651	7.41980	0.37	0.7158

### Example: AIDS Clinical Trials Group (ACTG) Protocol 320

- randomized, double-blind, placebo-controlled clinical trial
- eligibility criteria
  - HIV-infected adults
  - CD4 counts  $\leq 200$  and at least 3 months of prior zidovudine therapy
- two treatment groups
  - 3-drug regimen: indinavir, lamivudine, and either zidovudine or stavudine
  - 2-drug regimen: zidovudine and lamivudine
- 1156 patients randomized

- SAS procedures for describing binary, ordinal, and nominal data
  - proc freq
  - proc chart (or gchart)
  - proc tabulate
- SAS procedures for describing quantitative data
  - proc means
  - proc univariate
    - \* most thorough description
    - \* also does one-sample t-tests
  - proc tabulate

- patients stratified according to their CD4 count at study entry
  - $\leq 50$  cells/mm<sup>3</sup>
  - 50-200 cells/mm<sup>3</sup>
- primary endpoint: occurrence of an AIDS-defining event or death

## CD4 and RNA data from ACTG 320

- blood specimens collected at study entry and at weeks 4, 8, 24, and 40 during follow-up for analysis of CD4 counts and viral load
- patients included in the present analysis
  - 198 patients who were randomly selected for a virology substudy
  - ACTG 320 dataset available for purchase from National Technical Information Service includes clinical endpoints and CD4 data for all patients but viral load data only on these 198.

## The baseline data

- data collected one time on each patient at the time of entry into the study
- documentation that came with purchased data said this file had been written using the following SAS code

```
data _null_;
  file "basedata.dat" lrecl=36;
  set a.basedata;
  put
  @1 pidnum
  @7 sex
  @9 raceth
  @11 ivdrug
  @13 hemophil
  @15 karnof
  @19 avecd4
  @25 priorzdv
  @28 age
  ;
  run;
```

## Further documentation

5) BASEDATA.SSD01

\*\*\*\*\*

#	Variable	Type	Len	Pos	Label
8	AGE	Num	8	56	Age (years)
6	AVECD4	Num	8	40	Baseline CD4 Count
4	HEMOPHIL	Num	8	24	Hemophiliac? (1=yes, 0=no)
3	IVDRUG	Num	8	16	IV drug history
5	KARNOF	Num	8	32	Karnofsky score
9	PIDNUM	Num	8	64	
7	PRIORZDV	Num	8	48	Months prior ZDV
2	RACETH	Num	8	8	Race/ethnicity
1	SEX	Num	8	0	Sex (1=male, 2=female)

6=Other/unknown  
SEX Sex (1=Male, 2=Female)

HEMOPHIL Hemophiliac? (1=Yes, 0=No)

IVDRUG IV drug history (1=Never,2=Currently,3=Previously)

KARNOF Karnofsky Performance Scale  
coding: 100 = Normal; no complaint; no evidence of disease  
90 = Normal activity possible; minor signs/symptoms of disease  
80 = Normal activity with effort; some signs/symptoms of disease'  
70 = Cares for self; normal activity/active work not possible'

RACETH Race/ethnicity  
coding: 1=White Non-Hispanic  
2=Black Non-Hispanic  
3=Hispanic (Regardless of Race)  
4=Asian, Pacific Islander  
5=American Indian, Alaskan Native

● so I read it in and did descriptive statistics on each variable using following code

```
options linesize = 72 ;

proc format ;
    value sexfmt 1 = 'M' 2 = 'F' ;
    value racefmt 1 = 'W' 2 = 'B' 3 = 'H' 4 = 'A' 5 = 'NA' 6 = 'O' ;
    value yesfmt 1 = 'Y' 0 = 'N' ;
    value drugfmt 1 = 'Never' 2 = 'Current' 3 = 'Prev' ;
run ;

data base320 ;
infile '/group/markers/actg320/data/BASEDATA.DAT' ;
input
@1 pidnum
@7 sex
@9 raceth
@11 ivdrug
@13 hemophil
@15 karnof
@19 avecd4
@25 priorzdv
@28 age
;
format sex sexfmt. raceth racefmt. ivdrug drugfmt. hemophil yesfmt. ;

run;

proc print data = base320 (obs=25) ;
run ;

proc freq ;
```

```
tables sex raceth ivdrug hemophil ;
run ;

proc univariate plot ;
var age ; * if "var" statement is omitted, automatically does all ;
* numeric variables ;
run ;

proc means n mean std median min max ;
var age karnof avecd4 priorzdv ;
run ;
```

Obs	pidnum	sex	raceth	ivdrug	hemophil	karnof	avecd4	priorzdv	age
1	10333	M	W	Never	N	90	95.0	80	39.9425
2	10350	M	W	Never	N	90	185.0	60	51.3347
3	10491	M	W	Prev	N	90	180.0	60	63.4141
4	10719	M	W	Never	N	100	52.0	43	30.5106
5	11158	F	B	Never	N	90	188.5	45	40.4435
6	11656	M	W	Never	N	90	120.0	48	48.5339
7	11760	M	W	Prev	N	80	60.0	20	43.2033
8	11767	M	H	Never	N	90	85.0	6	43.3292
9	11796	M	W	Never	N	80	10.0	6	44.7036
10	11806	F	W	Never	N	80	40.0	8	49.2758
11	11851	M	H	Never	N	100	19.5	15	38.5681
12	11884	M	W	Never	N	90	12.0	9	45.8042
13	11898	M	H	Never	N	100	151.0	60	30.6557
14	11899	M	W	Never	N	100	75.0	7	25.6427
15	11900	M	W	Prev	N	100	97.0	14	33.6400
16	11901	M	H	Never	N	90	31.0	85	43.0554
17	11902	M	W	Never	N	90	51.5	11	34.2259
18	11904	M	W	Never	N	90	3.0	18	38.3272
19	11905	M	H	Never	N	100	112.5	36	32.0958
20	11921	M	W	Never	N	80	15.0	71	40.7474
21	11970	M	W	Never	N	100	112.0	26	31.8905
22	12144	M	H	Never	N	90	128.0	24	28.6160
23	12283	F	W	Never	N	100	161.5	7	30.2916
24	12284	M	W	Never	N	80	12.0	24	39.5209
25	12286	M	W	Never	N	90	153.5	8	42.7296

The FREQ Procedure

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
M	735	63.64	735	63.64
F	154	13.33	889	76.97
*	266	23.03	1155	100.00

  

raceth	Frequency	Percent	Cumulative Frequency	Cumulative Percent
W	506	43.81	506	43.81
B	240	20.78	746	64.59
H	123	10.65	869	75.24
A	13	1.13	882	76.36
NA	7	0.61	889	76.97
**	266	23.03	1155	100.00

  

ivdrug	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Never	849	73.51	849	73.51
Current	90	7.79	939	81.30
Prev	211	18.27	1150	99.57
4	1	0.09	1151	99.65
5	4	0.35	1155	100.00

hemophil	Frequency	Percent	Frequency	Percent
N	872	75.50	872	75.50
Y	231	20.00	1103	95.50
2	1	0.09	1104	95.58
3	51	4.42	1155	100.00

## Problem was in original writing of file

```

4301352 3 1 0 100 150 10
20.698151951
4401332 3 1 0 90 8 9
16.703627652
4401442 2 1 0 80 0 6
16.227241615
4604642 1 1 0 100 78 10
28.287474333
4705212 1 1 0 90 260 6
20.199863107
4802571 1 1 0 90 26.333633333
33.3908282
5001752 3 1 0 90 22.5 8
26.302532512
5017582 3 1 0 90 93 30
32.5229295

```

## Solution and correct output

- show length of pidnum variable
- correct data in datafile for records with one field written on top of another

```

input
@1 pidnum 1-6
@7 sex
  raceth
  ivdrug
  hemophil
  karnof
  avecd4
  priorzdv
  age
/* @1 pidnum
@7 sex
@9 raceth
@11 ivdrug
@13 hemophil
@15 karnof
@19 avecd4
@25 priorzdv
@28 age */
;
format sex sexfmt. raceth racefmt. ivdrug drugfmt. hemophil yesfmt. ;
run ;
.
.
.
proc print data = base320 ;
where not (raceth in (1,2,3,4,5,6)) ;
run ;

proc print data = base320 ;
where pidnum in (480257, 500175, 501758) ;
run ;

```

The FREQ Procedure

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
M	956	82.70	956	82.70
F	200	17.30	1156	100.00

  

raceth	Frequency	Percent	Cumulative Frequency	Cumulative Percent
W	598	51.73	598	51.73
B	328	28.37	926	80.10
H	205	17.73	1131	97.84
A	14	1.21	1145	99.05
NA	11	0.95	1156	100.00

  

ivdrug	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Never	972	84.08	972	84.08
Current	4	0.35	976	84.43
Prev	180	15.57	1156	100.00

  

hemophil	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	1120	96.89	1120	96.89
Y	36	3.11	1156	100.00

The UNIVARIATE Procedure  
Variable: age

Moments

N	1156	Sum Weights	1156
Mean	39.1568485	Sum Observations	45265.3169
Std Deviation	8.78515155	Variance	77.1788878
Skewness	0.6049763	Kurtosis	0.67779963
Uncorrected SS	1861588.77	Corrected SS	89141.6154
Coeff Variation	22.4357983	Std Error Mean	0.25838681

Basic Statistical Measures

Location		Variability	
Mean	39.15685	Std Deviation	8.78515
Median	38.16153	Variance	77.17889
Mode	38.10541	Range	57.95756
		Interquartile Range	11.21834

Tests for Location: Mu0=0

Test	-Statistic-	-----p Value-----	
Student's t	t 151.5435	Pr >  t	<.0001
Sign	M 578	Pr >=  M	<.0001
Signed Rank	S 334373	Pr >=  S	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	73.9302
99%	63.7892
95%	56.2136
90%	50.8392
75% Q3	44.3395
50% Median	38.1615
25% Q1	33.1211
10%	29.3169
5%	26.4476
1%	20.6982
0% Min	15.9726

Extreme Observations

-----Lowest-----		-----Highest-----	
Value	Obs	Value	Obs
15.9726	851	67.0308	68
16.2272	856	67.9973	41

The UNIVARIATE Procedure  
Variable: age

Extreme Observations

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

Value	Obs	Value	Obs
16.6078	849	71.5756	982
16.7036	855	73.2238	963
16.8734	853	73.9302	58

Histogram

	#	Boxplot
72.5+*	3	0
.*	6	0
***	16	0
*****	40	
*****	63	
*****	137	
*****	213	+-----+
*****	270	*--+-*
*****	266	+-----+
*****	108	
*****	26	
17.5+**	8	0

\* may represent up to 6 counts

