Computer Handout One

```
bash-3.00$ ls

bash-3.00$ curl http://fisher.utstat.toronto.edu/~brunner/2453f08/code_n_data/
lecture/senic.data > senic.data

  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  5989  100  5989    0     0   358k      0 --:--:-- --:--:-- --:--:-- 1848k

bash-3.00$  ls
senic.data
bash-3.00$ less senic.data
    1  7.13 55.7 4.1  9.0  39.6 279 2 4 207 241 60.0
    2  8.82 58.2 1.6  3.8  51.7  80 2 2  51   0 40.0
    3  8.34 9999 2.7  8.1  74.0 107 2 3  82  54 20.0
    4  8.95 53.7 5.6 18.9 122.8 147 2 4  53 148 40.0
    5 11.20 56.5 999 34.5  88.9 180 2 1 134 151 40.0


                    ... Skipping ...

  109 11.80 53.8 5.7  9.1 116.9 571 1 2 441 469 62.9
  110  9.50 49.3 5.8 42.0  70.9  98 2 3  68  46 22.9
  111  7.70 56.9 4.4 12.2  67.9 129 2 4  85 136 62.9
  112 17.94 56.2 5.9 26.4  91.8 835 1 1 791 407 62.9
  113  9.41 59.5 3.1 20.6  91.7  29 2 3  20  22 22.9
(END)
```

Spacebar for another page, q for quit

```
bash-3.00$ emacs senic0.sas
```

```
/*******  senic0.sas   ***********/
options linesize=79 noovp formdlim='_';

data simple;
     infile 'senic.data';
     input id stay age infrisk culratio xratio nbeds medschl
              region census nurses service;

proc freq;
     tables id -- service;    /*  Single dash only works with numbered
                                  lists, like item1-item50    */
```

```
bash-3.00$ ls
senic.data  senic0.sas
bash-3.00$ sas senic0
bash-3.00$ ls
senic.data  senic0.log  senic0.lst  senic0.sas
bash-3.00$ cat senic0.log
```

```
1                                                          The SAS System

NOTE: Copyright (c) 2002-2003 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) 9.1 (TS1M0)
      Licensed to UNIVERSITY OF TORONTO/COMPUTING & COMMUNICATIONS, Site 0008987001.
NOTE: This session is executing on the SunOS 5.10 platform.



You are running SAS 9. Some SAS 8 files will be automatically converted
by the V9 engine; others are incompatible.  Please see
http://support.sas.com/rnd/migration/planning/platform/64bit.html

PROC MIGRATE will preserve current SAS file attributes and is
recommended for converting all your SAS libraries from any
SAS 8 release to SAS 9.  For details and examples, please see
http://support.sas.com/rnd/migration/index.html


This message is contained in the SAS news file, and is presented upon
initialization.  Edit the file "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.




NOTE: SAS initialization used:
      real time           0.13 seconds
      cpu time            0.11 seconds

1          /*******  senic0.sas   ***********/
2          options linesize=79 noovp formdlim='_';
3
4          data simple;
5               infile 'senic.data';
6               input id stay age infrisk culratio xratio nbeds medschl
7                        region census nurses service;
8

NOTE: The infile 'senic.data' is:
      File Name=/u/brunner/2453f08/show/senic.data,
      Owner Name=brunner,Group Name=UNKNOWN,
      Access Permission=rw-r--r--,
      File Size (bytes)=5989

NOTE: 113 records were read from the infile 'senic.data'.
      The minimum record length was 52.
      The maximum record length was 52.
NOTE: The data set WORK.SIMPLE has 113 observations and 12 variables.
NOTE: DATA statement used (Total process time):
      real time           0.04 seconds
      cpu time            0.02 seconds
```

```
9          proc freq;
10              tables id -- service;    /*  Single dash only works with
10       ! numbered
11                                           lists, like item1-item50    */
12

2                                The SAS System

NOTE: There were 113 observations read from the data set WORK.SIMPLE.
NOTE: The PROCEDURE FREQ printed pages 1-20.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.15 seconds
      cpu time            0.12 seconds


NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           0.35 seconds
      cpu time            0.26 seconds

bash-3.00$  less senic0.lst

It starts like this ...
```

```
                                The SAS System                                1

                             The FREQ Procedure

                                                 Cumulative    Cumulative
           id    Frequency     Percent     Frequency      Percent
          ----------------------------------------------------------
            1            1        0.88             1         0.88
            2            1        0.88             2         1.77
            3            1        0.88             3         2.65
            4            1        0.88             4         3.54
            5            1        0.88             5         4.42
            6            1        0.88             6         5.31
            7            1        0.88             7         6.19
            8            1        0.88             8         7.08
            9            1        0.88             9         7.96
           10            1        0.88            10         8.85
```

Displaying frequency distributions for all the variables. We will skip around.
Lowest average age is 38.8. Continuing ...

age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
58	1	0.88	100	88.50
58.2	2	1.77	102	90.27
59	1	0.88	103	91.15
59.5	1	0.88	104	92.04
59.6	1	0.88	105	92.92
59.9	1	0.88	106	93.81
60.9	1	0.88	107	94.69
61.1	1	0.88	108	95.58
62.2	1	0.88	109	96.46
63.9	1	0.88	110	97.35
64.1	1	0.88	111	98.23
65.9	1	0.88	112	99.12
9999	1	0.88	113	100.00

That's pretty old. Infection risk ...

infrisk	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1.3	2	1.77	2	1.77
1.4	1	0.88	3	2.65
1.6	1	0.88	4	3.54
1.7	1	0.88	5	4.42
1.8	1	0.88	6	5.31
2	2	1.77	8	7.08
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
6.6	1	0.88	109	96.46
7.6	1	0.88	110	97.35
7.7	1	0.88	111	98.23
7.8	1	0.88	112	99.12
999	1	0.88	113	100.00

Now take a look at the categorical variables.

medschl	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	17	15.04	17	15.04
2	96	84.96	113	100.00

region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	29	25.66	29	25.66
2	32	28.32	61	53.98
3	36	31.86	97	85.84
4	16	14.16	113	100.00

```
bash-3.00$ ls
senic.data  senic0.log  senic0.lst  senic0.sas
bash-3.00$ cp senic0.sas senic1.sas
bash-3.00$ emacs senic1.sas

/*******  senic1.sas  Read and describe SENIC data  ***********/
options linesize=79 noovp formdlim='_';
title 'Read and Describe SENIC data';

proc format;  /* value labels used in data step below */
     value yesnofmt 1 = 'Yes'   2 = 'No' ;
     value regfmt  1 =  'Northeast'
                   2 =  'North Central'
                   3 =  'South'
                   4 =  'West' ;
     value acatfmt 1 = '53 & under' 2 = 'over 53';

data better;
     infile 'senic.data';
     input id stay age infrisk culratio xratio nbeds medschl
              region census nurses service;

      /*** SAS doesn't like numeric missing value codes.  A period  .  is
                       best for missing. However ...  ***/

      if stay eq 9999 then stay = .  ;
      if age eq 9999 then age = .  ;
      if infrisk = 999 then infrisk = .;
      if culratio = 9999 then culratio = .;
      if xratio = 9999 then xratio = .;
      if nurses eq 0 then nurses = .  ;

     label id       = 'Hospital identification number'
           stay     = 'Av length of hospital stay, in days'
           age      = 'Average patient age'
           infrisk  = 'Prob of acquiring infection in hospital'
           culratio = '# cultures  / # no hosp acq infect'
           xratio   = '# x-rays / # no signs of pneumonia'
           nbeds    = 'Average # beds during study period'
           medschl  = 'Medical school affiliation'
           region   = 'Region of country (usa)'
           census   = 'Aver # patients in hospital per day'
           nurses   = 'Aver # nurses during study period'
           service  = '% of 35 potential facil. & services' ;
        /* Associating variables with their value labels */
     format medschl yesnofmt.;
     format region  regfmt.;


    /*************** Create some new variables  ***************/

     if 0<age<=53 then agecat=1;
     else if age>53 then agecat=2;
     label   agecat = 'av patient age category';
     format agecat acatfmt.;
```

```
   /*  Compute ad hoc index of hospital quality  */
     quality=(2*service+nurses+nbeds+10*culratio
                 +10*xratio-2*stay)/medschl;
     if (region eq 3) then quality=quality-100;
     label quality = "Jerry's bogus hospital quality index";

     /* Dummy variables (There are no missing values) */
     if region = 1 then r1=1; else r1=0;
     if region = 2 then r2=1; else r2=0;
     if region = 3 then r3=1; else r3=0;
     if region = 4 then r4=1; else r4=0;

     if medschl = 2 then mschool = 0; else mschool = medschl;
     /* mschool is an indicator for medical school = yes */

   /* reg1-reg3 & ms1 are effect coded dummy vars */
     ms1 = medschl; if medschl = 2 then ms1 = -1;
     if region = 1 then reg1 = 1;
        else if region=4 then reg1 = -1;
        else reg1 = 0;
     if region = 2 then reg2 = 1;
        else if region=4 then reg2 = -1;
        else reg2 = 0;
     if region = 3 then reg3 = 1;
        else if region=4 then reg3 = -1;
        else reg3 = 0;
     /*  Interaction terms */
     mr1 = ms1 * reg1; mr2 = ms1 * reg2 ; mr3 = ms1 * reg3;

proc means;
     title2 'Basic Descriptive Stats for Quantitative Vars';
     var stay age infrisk culratio xratio nbeds census nurses service
         quality;

proc univariate normal plot;
     title2 'More detail for infrisk';
     var infrisk;

proc freq;
     title2 'Frequency distributions of Categorical Variables';
     tables region medschl agecat;

proc freq;
     title2 'Check Dummy variables';
     tables medschl * (mschool ms1) / norow nocol nopercent missprint;
     tables region * (r1-r4 reg1-reg3)
                                   / norow nocol nopercent missprint;



bash-3.00$ ls
senic.data  senic0.log  senic0.lst  senic0.sas  senic1.sas
bash-3.00$ sas senic1
bash-3.00$ ls
senic.data  senic0.lst  senic1.log  senic1.sas
senic0.log  senic0.sas  senic1.lst
bash-3.00$ less senic1.log
bash-3.00$ cat senic1.lst
```

Basic Descriptive Stats for Quantitative Vars

The MEANS Procedure

Variable	Label	N	Mean
stay	Av length of hospital stay, in days	113	9.6483186
age	Average patient age	112	53.1991071
infrisk	Prob of acquiring infection in hospital	112	4.3428571
culratio	# cultures / # no hosp acq infect	112	15.6750000
xratio	# x-rays / # no signs of pneumonia	112	81.6535714
nbeds	Average # beds during study period	113	252.1769912
census	Aver # patients in hospital per day	113	191.3716814
nurses	Aver # nurses during study period	112	174.3303571
service	% of 35 potential facil. & services	113	43.1548673
quality	Jerry's bogus hospital quality index	111	863.1233333

Variable	Label	Std Dev	Minimum
stay	Av length of hospital stay, in days	1.9114560	6.7000000
age	Average patient age	4.4679942	38.8000000
infrisk	Prob of acquiring infection in hospital	1.3408156	1.3000000
culratio	# cultures / # no hosp acq infect	10.2283522	1.6000000
xratio	# x-rays / # no signs of pneumonia	19.4521632	39.6000000
nbeds	Average # beds during study period	192.8451558	29.0000000
census	Aver # patients in hospital per day	153.7595639	20.0000000
nurses	Aver # nurses during study period	139.4128866	14.0000000
service	% of 35 potential facil. & services	15.2001879	5.7000000
quality	Jerry's bogus hospital quality index	565.9398127	188.7300000

Variable	Label	Maximum
stay	Av length of hospital stay, in days	19.5600000
age	Average patient age	65.9000000
infrisk	Prob of acquiring infection in hospital	7.8000000
culratio	# cultures / # no hosp acq infect	60.5000000
xratio	# x-rays / # no signs of pneumonia	133.5000000
nbeds	Average # beds during study period	835.0000000
census	Aver # patients in hospital per day	791.0000000
nurses	Aver # nurses during study period	656.0000000
service	% of 35 potential facil. & services	80.0000000
quality	Jerry's bogus hospital quality index	3066.86

The UNIVARIATE Procedure
Variable: infrisk (Prob of acquiring infection in hospital)

Moments

N	112	Sum Weights	112
Mean	4.34285714	Sum Observations	486.4
Std Deviation	1.34081556	Variance	1.79778636
Skewness	-0.1027819	Kurtosis	0.19954989
Uncorrected SS	2311.92	Corrected SS	199.554286
Coeff Variation	30.8740424	Std Error Mean	0.12669516

Basic Statistical Measures

Location		Variability	
Mean	4.342857	Std Deviation	1.34082
Median	4.400000	Variance	1.79779
Mode	4.300000	Range	6.50000
		Interquartile Range	1.60000

NOTE: The mode displayed is the smallest of 2 modes with a count of 7.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	34.278	Pr > |t|	<.0001
Sign	M	56	Pr >= |M|	<.0001
Signed Rank	S	3164	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.981972	Pr < W	0.1357
Kolmogorov-Smirnov	D	0.097777	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.127638	Pr > W-Sq	0.0475
Anderson-Darling	A-Sq	0.702387	Pr > A-Sq	0.0685

Quantiles (Definition 5)

Quantile	Estimate
100% Max	7.8
99%	7.7
95%	6.4
90%	5.8
75% Q3	5.2
50% Median	4.4
25% Q1	3.6
10%	2.6
5%	1.8

The UNIVARIATE Procedure
Variable: infrisk (Prob of acquiring infection in hospital)

Quantiles (Definition 5)

Quantile	Estimate
1%	1.3
0% Min	1.3

Extreme Observations

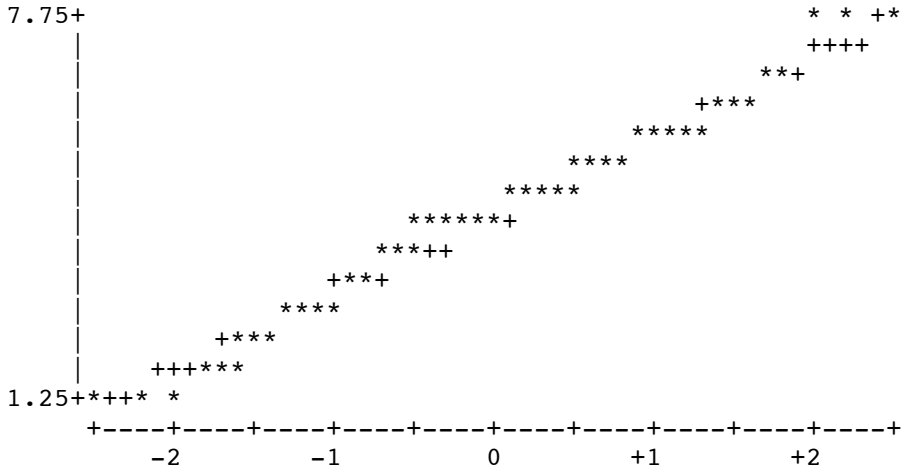
----Lowest----		----Highest---	
Value	Obs	Value	Obs
1.3	93	6.5	47
1.3	40	6.6	104
1.4	107	7.6	53
1.6	2	7.7	13
1.7	85	7.8	54

Missing Values

Missing Value	Count	-----Percent Of----- All Obs	Missing Obs
.	1	0.88	100.00

```
 Stem Leaf                     #             Boxplot
    7 678                      3                0
    7                                           |
    6 56                       2                |
    6 12334                    5                |
    5 555566677889            12                |
    5 0000112233344           13             +-----+
    4 5555555666778888999     19             |     |
    4 0111122222333333344444  22             *--+--*
    3 5577778999              10             +-----+
    3 011244                   6                |
    2 5677899999              10                |
    2 0013                     4                |
    1 678                      3                |
    1 334                      3                |
      ----+----+----+----+--
```

The UNIVARIATE Procedure
Variable: infrisk (Prob of acquiring infection in hospital)



Frequency distributions of Categorical Variables

The FREQ Procedure

Region of country (usa)

region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Northeast	29	25.66	29	25.66
North Central	32	28.32	61	53.98
South	36	31.86	97	85.84
West	16	14.16	113	100.00

Medical school affiliation

medschl	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Yes	17	15.04	17	15.04
No	96	84.96	113	100.00

av patient age category

agecat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
53 & under	56	50.00	56	50.00
over 53	56	50.00	112	100.00

Frequency Missing = 1

Check Dummy variables

The FREQ Procedure

Table of medschl by mschool

medschl(Medical school affiliation)

Frequency | mschool	0	1	Total
Yes	0	17	17
No	96	0	96
Total	96	17	113

Table of medschl by ms1

medschl(Medical school affiliation)

Frequency	ms1 -1	1	Total
Yes	0	17	17
No	96	0	96
Total	96	17	113

Table of region by r1

region(Region of country (usa))

Frequency	r1 0	1	Total
Northeast	0	29	29
North Central	32	0	32
South	36	0	36
West	16	0	16
Total	84	29	113

Read and Describe SENIC data 7
Check Dummy variables

The FREQ Procedure

Table of region by r2

region(Region of country (usa))

Frequency	r2 0	1	Total
Northeast	29	0	29
North Central	0	32	32
South	36	0	36
West	16	0	16
Total	81	32	113

Table of region by r3

region(Region of country (usa))

Frequency	r3 0	r3 1	Total
Northeast	29	0	29
North Central	32	0	32
South	0	36	36
West	16	0	16
Total	77	36	113

Table of region by r4

region(Region of country (usa))

Frequency	r4 0	r4 1	Total
Northeast	29	0	29
North Central	32	0	32
South	36	0	36
West	0	16	16
Total	97	16	113

The FREQ Procedure

Table of region by reg1

region(Region of country (usa)) reg1

Frequency	-1	0	1	Total
Northeast	0	0	29	29
North Central	0	32	0	32
South	0	36	0	36
West	16	0	0	16
Total	16	68	29	113

```
                    Table of region by reg2

region(Region of country (usa))     reg2

Frequency     |      -1|       0|       1|  Total
--------------+--------+--------+--------+
Northeast     |       0 |     29 |       0 |     29
--------------+--------+--------+--------+
North Central |       0 |       0 |     32 |     32
--------------+--------+--------+--------+
South         |       0 |     36 |       0 |     36
--------------+--------+--------+--------+
West          |      16 |       0 |       0 |     16
--------------+--------+--------+--------+
Total               16       65       32      113


                    Table of region by reg3

region(Region of country (usa))     reg3

Frequency     |      -1|       0|       1|  Total
--------------+--------+--------+--------+
Northeast     |       0 |     29 |       0 |     29
--------------+--------+--------+--------+
North Central |       0 |     32 |       0 |     32
--------------+--------+--------+--------+
South         |       0 |       0 |     36 |     36
--------------+--------+--------+--------+
West          |      16 |       0 |       0 |     16
--------------+--------+--------+--------+
Total               16       61       36      113

bash-3.00$
```

Make a pure data definition file.

```
bash-3.00$ cp senic1.sas senicdef.sas
bash-3.00$ emacs senicdef.sas

/*******  senicdef.sas  Just read the SENIC data  ***********/
options linesize=79 noovp formdlim='_';
title 'Study of the Efficacy of Nosocomial Infection Control (SENIC)';

proc format;  /* value labels used in data step below */
     value yesnofmt 1 = 'Yes'   2 = 'No' ;
     value regfmt  1 =  'Northeast'
                   2 =  'North Central'
                   3 =  'South'
                   4 =  'West' ;
     value acatfmt 1 = '53 & under' 2 = 'over 53';

data better;
     infile 'senic.data';
     input id stay age infrisk culratio xratio nbeds medschl
              region census nurses service;
```

```
      /*** SAS doesn't like numeric missing value codes.  A period  .  is
                       best for missing. However ...  ***/

       if stay eq 9999 then stay = .  ;
       if age eq 9999 then age = .  ;
       if infrisk = 999 then infrisk = .;
       if culratio = 9999 then culratio = .;
       if xratio = 9999 then xratio = .;
       if nurses eq 0 then nurses = .  ;

      label id       = 'Hospital identification number'
            stay     = 'Av length of hospital stay, in days'
            age      = 'Average patient age'
            infrisk  = 'Prob of acquiring infection in hospital'
            culratio = '# cultures  / # no hosp acq infect'
            xratio   = '# x-rays / # no signs of pneumonia'
            nbeds    = 'Average # beds during study period'
            medschl  = 'Medical school affiliation'
            region   = 'Region of country (usa)'
            census   = 'Aver # patients in hospital per day'
            nurses   = 'Aver # nurses during study period'
            service  = '% of 35 potential facil. & services' ;
         /* Associating variables with their value labels */
       format medschl yesnofmt.;
       format region  regfmt.;


     /*************** Create some new variables  ****************/

       if 0<age<=53 then agecat=1;
       else if age>53 then agecat=2;
       label   agecat = 'av patient age category';
       format agecat acatfmt.;

    /*  Compute ad hoc index of hospital quality  */
       quality=(2*service+nurses+nbeds+10*culratio
                   +10*xratio-2*stay)/medschl;
       if (region eq 3) then quality=quality-100;
       label quality = "Jerry's bogus hospital quality index";

/* Dummy variables (There are no missing values) */
       if region = 1 then r1=1; else r1=0;
       if region = 2 then r2=1; else r2=0;
       if region = 3 then r3=1; else r3=0;
       if region = 4 then r4=1; else r4=0;

       if medschl = 2 then mschool = 0; else mschool = medschl;
       /* mschool is an indicator for medical school = yes */
```

```
/* reg1-reg3 & ms1 are effect coded dummy vars */
     ms1 = medschl; if medschl = 2 then ms1 = -1;
     if region = 1 then reg1 = 1;
        else if region=4 then reg1 = -1;
        else reg1 = 0;
     if region = 2 then reg2 = 1;
        else if region=4 then reg2 = -1;
        else reg2 = 0;
     if region = 3 then reg3 = 1;
        else if region=4 then reg3 = -1;
        else reg3 = 0;
     /*  Interaction terms */
     mr1 = ms1 * reg1; mr2 = ms1 * reg2 ; mr3 = ms1 * reg3;

bash-3.00$ cp senic1.sas senic1b.sas

bash-3.00$ emacs senic1b.sas




/*******  senic1b.sas  Describe SENIC data  ***********/
/* senic1b.sas */
%include 'senicdef.sas'; /* Effectively, Copy the file senicdef.sas to here */

proc means;
     title2 'Basic Descriptive Stats for Quantitative Vars';
     var stay age infrisk culratio xratio nbeds census nurses service
         quality;

proc univariate normal plot;
     title2 'More detail for infrisk';
     var infrisk;

proc freq;
     title2 'Frequency distributions of Categorical Variables';
     tables region medschl agecat;

proc freq;
     title2 'Check Dummy variables';
     tables medschl * (mschool ms1) / norow nocol nopercent missprint;
     tables region * (r1-r3 rg1-rg3 reg1-reg3)
                                    / norow nocol nopercent missprint;
bash-3.00$
```