Introduction to SAS: The SENIC data

```
/dos/brunner/2201/senic/reg > ls
2201senic1.sas     2201senicdef.sas    senic.raw
2201senic2.sas     2201senicreg.sas
/dos/brunner/2201/senic/reg > sas 2201senic1
/dos/brunner/2201/senic/reg > ls
2201senic1.log     2201senic1.sas      2201senicdef.sas    senic.raw
2201senic1.lst     2201senic2.sas      2201senicreg.sas

/dos/brunner/2201/senic/reg > head senic.raw
     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  52 40.0
     3   8.34 56.9 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 5.7  34.5   88.9 180 2 1 134 151 40.0
     6   9.76 50.9 5.1  21.9   97.0 150 2 2 147 106 40.0
     7   9.68 57.8 4.6  16.7   79.0 186 2 3 151 129 40.0
     8  11.18 45.7 5.4  60.5   85.8 640 1 2 399 360 60.0
     9   8.67 48.2 4.3  24.4   90.8 182 2 3 130 118 40.0
    10   8.84 56.3 6.3  29.6   82.6  85 2 1  59  66 40.0
```

```sas
/* 2201senic1.sas */
title 'Read and Describe SENIC data: Basic';
options linesize=79 pagesize=35 noovp formdlim='_';

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

proc means n mean stddev;
    title2  'Describe continuous variables';
    var stay -- nbeds census nurses service;
proc freq;
    title2  'Describe categorical variables';
    tables medschl region;
```

```
/dos/brunner/2201/senic/reg > cat 2201senic1.log
1                                                  The SAS System
00:00 Thursday, February 16, 2006

NOTE: Copyright (c) 1999-2001 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.2 (TS2M0)
      Licensed to UNIVERSITY OF TORONTO/COMPUTING & COMMUNICATIONS, Site 0008987001.
NOTE: This session is executing on the SunOS 5.9 platform.


This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "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             0.77 seconds
      cpu time              0.05 seconds
```

```
1           /* 2201senic1.sas */
2           title 'Read and Describe SENIC data: Basic';
3           options linesize=79 pagesize=35 noovp formdlim='_';
4
5           data  simple;
6                 infile 'senic.raw';
7                 input   id stay age infrisk culratio xratio nbeds medschl
8                         region census nurses service;
9
NOTE: The infile 'senic.raw' is:
      File Name=/u/brunner/2201/senic/reg/senic.raw,
      Owner Name=brunner,Group Name=dos,
      Access Permission=rw-r--r--,

2                               The SAS System
                                          00:00 Thursday, February 16, 2006

      File Size (bytes)=5989

NOTE: 113 records were read from the infile 'senic.raw'.
      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:
      real time            0.28 seconds
      cpu time             0.04 seconds


10          proc means n mean stddev;
11               title2  'Describe continuous variables';
12               var stay -- nbeds census nurses service;

NOTE: There were 113 observations read from the data set WORK.SIMPLE.
NOTE: The PROCEDURE MEANS printed page 1.
NOTE: PROCEDURE MEANS used:
      real time            0.34 seconds
      cpu time             0.02 seconds


13          proc freq;
14               title2  'Describe categorical variables';
15               tables medschl region;
16
NOTE: There were 113 observations read from the data set WORK.SIMPLE.
NOTE: The PROCEDURE FREQ printed page 2.
NOTE: PROCEDURE FREQ used:
      real time            0.09 seconds
      cpu time             0.01 seconds

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time            1.71 seconds
      cpu time             0.12 seconds
```

```
/dos/brunner/2201/senic/reg >
/dos/brunner/2201/senic/reg > cat 2201senic1.lst
```

                                                    00:00 Thursday, February 16, 2006

                              The MEANS Procedure

              Variable   N   Mean   Std Dev
              ----
              stay   113   9.6483186   1.9114560
              age   113   53.2318584   4.4616074
              infrisk   113   4.3548673   1.3409080
              culratio   113   15.6840708   10.1830441
              xratio   113   81.6300885   19.3667373
              nbeds   113   252.1769912   192.8451558
              census   113   191.3716814   153.7595639
              nurses   113   173.2477876   139.2653897
              service   113   43.1548673   15.2001879

                                                    00:00 Thursday, February 16, 2006

                              The FREQ Procedure

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

                                             Cumulative   Cumulative
            region   Frequency   Percent   Frequency   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
```

```
/* 2201senicdef.sas */
title 'SENIC data';
options linesize=79 pagesize=35 noovp formdlim='_';

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.raw';
      input id stay age infrisk culratio xratio nbeds medschl
           region census nurses service;
     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.;

        /***** recodes, computes & ifs  *****/

     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 for region: South is the reference category */
     if region = 1 then r1=1; else r1=0;
     if region = 2 then r2=1; else r2=0;
     if region = 4 then r3=1; else r3=0;

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

```
/* 2201senic2.sas */
title2 'Descriptive Statistics and a Few Elementary Tests';
%include '2201senicdef.sas';
options pagesize=100;

proc means n mean stddev;
     title2  'Describe continuous variables';
     var stay -- nbeds census nurses service;
proc freq;
     title2  'Describe categorical variables';
     tables agecat medschl region;

PROC TTEST;
     title3 'Independent t-test with PROC TTEST';
     CLASS MEDSCHL;
     VAR INFRISK;

PROC GLM;
     title3 'One-way ANOVA and followups with PROC GLM';
     CLASS REGION;
     MODEL INFRISK=REGION;
     means region;
     MEANS REGION/ bon;

/* Labels can get in the way. Create a new SAS data set without labels */
data without;
     set better;
     label id       = ' '  stay     = ' '  age      = ' '  infrisk = ' '
           culratio = ' '  xratio   = ' '  nbeds    = ' '  medschl = ' '
           region   = ' '  census   = ' '  nurses   = ' '  service = ' ';

proc corr nosimple;
     var stay age nbeds census nurses xratio culratio;
     with infrisk;


/dos/brunner/2201/senic/reg > cat 2201senic2.lst
```

                        The MEANS Procedure

Variable	Label	N	Mean
stay	av length of hospital stay, in days	113	9.6483186
age	average patient age	113	53.2318584
infrisk	prob of acquiring infection in hospital	113	4.3548673
culratio	# cultures / # no hosp acq infect	113	15.6840708
xratio	# x-rays / # no signs of pneumonia	113	81.6300885
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	113	173.2477876
service	% of 35 potential facil. & services	113	43.1548673

```
--------------------------------------------------------------------------

     Variable    Label                                         Std Dev
     -----------------------------------------------------------------
     stay        av length of hospital stay, in days         1.9114560
     age         average patient age                         4.4616074
     infrisk     prob of acquiring infection in hospital     1.3409080
     culratio    # cultures  / # no hosp acq infect         10.1830441
     xratio      # x-rays / # no signs of pneumonia         19.3667373
     nbeds       average # beds during study period        192.8451558
     census      aver # patients in hospital per day       153.7595639
     nurses      aver # nurses during study period         139.2653897
     service     % of 35 potential facil. & services        15.2001879
     -----------------------------------------------------------------
```

SENIC data   2
Describe categorical variables
                           14:39 Sunday, February 19, 2006

The FREQ Procedure

av patient age category

```
                                     Cumulative    Cumulative
      agecat    Frequency    Percent   Frequency      Percent
      ----------------------------------------------------------
      53 & under       56      49.56          56        49.56
      Over 53          57      50.44         113       100.00
```

medical school affiliation

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

region of country (usa)

```
                                     Cumulative    Cumulative
       region    Frequency    Percent   Frequency      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
```

                        SENIC data   3
             Describe categorical variables
            Independent t-test with PROC TTEST
                              14:39 Sunday, February 19, 2006

                   The TTEST Procedure

                      Statistics

                    Lower CL   Upper CL Lower CL
Variable medschl   N   Mean   Mean   Mean   Std Dev   Std Dev

infrisk   Yes   17   4.5223   5.0941   5.6659   0.8283   1.1121
infrisk   No   96   3.9524   4.224   4.4955   1.1738   1.3403
infrisk   Diff (1-2)   0.1872   0.8702   1.5531   1.1579   1.3099

                      Statistics

                    Upper CL
        Variable medschl   Std Dev   Std Err   Minimum   Maximum

        infrisk   Yes   1.6926   0.2697   2.9   7.7
        infrisk   No   1.5622   0.1368   1.3   7.8
        infrisk   Diff (1-2)   1.5081   0.3447

                       T-Tests

      Variable   Method   Variances   DF   t Value   Pr > |t|

      infrisk   Pooled   Equal   111   2.52   0.0130
      infrisk   Satterthwaite   Unequal   25   2.88   0.0081

                 Equality of Variances

        Variable   Method   Num DF   Den DF   F Value   Pr > F

        infrisk   Folded F   95   16   1.45   0.4026

```
                         SENIC data                                    4
                  Describe categorical variables
                One-way ANOVA and followups with PROC GLM
                                     14:39 Sunday, February 19, 2006
```

```
                      The GLM Procedure

                  Class Level Information

        Class          Levels    Values

        region            4      North central Northeast South West


            Number of observations      113
```

```
                         SENIC data                                    5
                  Describe categorical variables
                One-way ANOVA and followups with PROC GLM
                                     14:39 Sunday, February 19, 2006
```

```
                      The GLM Procedure
```

Dependent Variable: infrisk   prob of acquiring infection in hospital

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	17.5750218	5.8583406	3.47	0.0186
Error	109	183.8048012	1.6862826		
Corrected Total	112	201.3798230			

R-Square	Coeff Var	Root MSE	infrisk Mean
0.087273	29.81881	1.298569	4.354867

Source	DF	Type I SS	Mean Square	F Value	Pr > F
region	3	17.57502176	5.85834059	3.47	0.0186

Source	DF	Type III SS	Mean Square	F Value	Pr > F
region	3	17.57502176	5.85834059	3.47	0.0186

                            SENIC data   6
                   Describe categorical variables
                One-way ANOVA and followups with PROC GLM
                                        14:39 Sunday, February 19, 2006

                          The GLM Procedure

        Level of   ----------infrisk----------
        region   N   Mean   Std Dev

        North central   32   4.39375000   1.33921920
        Northeast   29   4.90689655   1.27277285
        South   36   3.86388889   1.42751588
        West   16   4.38125000   0.87652248

                            SENIC data   7
                   Describe categorical variables
                One-way ANOVA and followups with PROC GLM
                                        14:39 Sunday, February 19, 2006

                          The GLM Procedure

             Bonferroni (Dunn) t Tests for infrisk

    NOTE: This test controls the Type I experimentwise error rate, but it
   generally has a higher Type II error rate than Tukey's for all pairwise
                             comparisons.

                 Alpha   0.05
                 Error Degrees of Freedom   109
                 Error Mean Square   1.686283
                 Critical Value of t   2.68726

       Comparisons significant at the 0.05 level are indicated by ***.

                              Difference
               region   Between   Simultaneous 95%
             Comparison   Means   Confidence Limits

Northeast   - North central   0.5131   -0.3815   1.4078
Northeast   - West   0.5256   -0.5611   1.6124
Northeast   - South   1.0430   0.1723   1.9137   ***
North central - Northeast   -0.5131   -1.4078   0.3815
North central - West   0.0125   -1.0560   1.0810
North central - South   0.5299   -0.3180   1.3777
West   - Northeast   -0.5256   -1.6124   0.5611
West   - North central   -0.0125   -1.0810   1.0560
West   - South   0.5174   -0.5311   1.5659
South   - Northeast   -1.0430   -1.9137   -0.1723   ***
South   - North central   -0.5299   -1.3777   0.3180
South   - West   -0.5174   -1.5659   0.5311

                              SENIC data   8
                     Describe categorical variables
                  One-way ANOVA and followups with PROC GLM
                                   14:39 Sunday, February 19, 2006

                          The CORR Procedure

   1 With Variables: infrisk
   7   Variables: stay   age   nbeds   census   nurses   xratio
                        culratio

                  Pearson Correlation Coefficients, N = 113
                      Prob > |r| under H0: Rho=0

                  stay   age   nbeds   census   nurses   xratio   culratio

infrisk   0.53344   0.00109   0.35976   0.38141   0.39398   0.45329   0.57914
               <.0001   0.9908   <.0001   <.0001   <.0001   <.0001   <.0001