STA429/1007 Handout 1: SENIC

```
appsrv01.srv> ls
senic0.sas  senic.dat
appsrv01.srv> sas senic0
appsrv01.srv> ls
senic0.log  senic0.lst  senic0.sas  senic.dat
appsrv01.srv>
appsrv01.srv> cat senic0.log
1                                                          The SAS System
13:59 Sunday, September 5, 2004

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 Linux 2.6.8.1-smp-athlon-bk 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           0.03 seconds
      cpu time            0.00 seconds

1          /* senic0.sas */
2          data  simple;
3                infile 'senic.dat';
4                input   id stay age infrisk culratio xratio nbeds medschl
5                        region census nurses service;

NOTE: The infile 'senic.dat' is:
      File Name=/homes/students/u0/stats/brunner/senic.dat,
      Owner Name=brunner,Group Name=stats,
      Access Permission=rw-r-----,
      File Size (bytes)=5989

NOTE: 113 records were read from the infile 'senic.dat'.
      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           0.00 seconds
      cpu time            0.02 seconds


6          proc freq;
7                tables _all_;
8
NOTE: There were 113 observations read from the data set WORK.SIMPLE.
NOTE: The PROCEDURE FREQ printed pages 1-20.
```

```
NOTE: PROCEDURE FREQ used:
      real time           0.02 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           0.11 seconds
      cpu time            0.03 seconds


appsrv01.srv> cp senic0.sas senic0.1.sas
appsrv01.srv> emacs senic0.1.sas

appsrv01.srv> cat senic0.1.sas

/* senic0.1.sas */
options linesize = 79;
data  simple;
      infile 'senic.dat';
      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 xratio eq 9999 then xratio = .  ;
      if culratio eq 9999 then culratio = .  ;
      if infrisk = 999 then infrisk  = .  ;
      if nbeds = 9 then nbeds  = .  ;
      if medschl = 9 then medschl  = .  ;
      if region = 9 then region  = .  ;
      if census = 9 then census  = .  ;
      if service = 9 then service  = .  ;
      if nurses eq (0 or .999) then nurses = .  ;

proc freq;
      tables _all_;


appsrv01.srv> sas senic0.1.lst
appsrv01.srv> less senic0.1.log
appsrv01.srv> less senic0.1.lst

                              The SAS System                                1
                                            14:21 Sunday, September 5, 2004

                             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
```

```
/*******  senicread.sas  just reads and labels SENIC data  *************
*                                                                      *
* Reading from the data file senic.raw, in which the variables are in  *
* fixed columns and missing data are represented by blanks.            *
*                                                                      *
************************************************************************/

title 'Study of the Effectiveness of Nosocomial Infection Control';
options linesize=79 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 senic;
     infile 'senic.raw' missover ; /*  in senic.raw, missing=blank  */
                                   /*  missover causes blanks to be missing */
     input
         #1  id       1-5
             stay     7-11
             age      13-16
             infrisk  18-20
             culratio 22-25
             xratio   27-31
             nbeds    33-35
             medschl  37
             region   39
             census   41-43
             nurses   45-47
             service  49-52 ;
     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';

/* Commented out

proc freq;
      tables _all_;



/****************** senicdescr.sas ***************/
/*            Descriptive stats on SENIC Data      */
/************************************************/

%include 'senicread.sas';  /* senicread.sas reads data, etc.  */
title2 'Descriptive Statistics';

proc freq;
     title3 'Frequency distributions of categorical variables';
     tables medschl region agecat;

proc means n mean std;
     title3 'Means and SDs of quantitative variables';
     var stay -- nbeds census nurses service;
     /* single dash only works with numbered lists, like item1-item50    */

proc univariate  plot normal ; /* Plots and a test for normality */
     title3 'Describe Quantitative Variables in More Detail' ;
     var stay -- nbeds census nurses service;
```

```
          Study of the Effectiveness of Nosocomial Infection Control          1
                            Descriptive Statistics
                 Frequency distributions of categorical variables
                                              14:58 Sunday, September 5, 2004

                              The FREQ Procedure

                           Medical school affiliation

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

```
/****************** basicsenic.sas ***************/
/*             Elementary tests on SENIC Data      */
/*************************************************/
%include 'senicread.sas';  /* senicread.sas reads data, etc.  */
title2 'Elementary tests on SENIC Data';

proc freq;
     title3 'Use proc freq to do crosstabs with chisquare test';
     tables region*medschl / nocol nopercent expected chisq;
proc ttest;
     title3 'T-test: Less risk at Hospitals with Med School Affiliation?';
     class medschl;
     var infrisk age ;
proc glm;
     title3 'One-way anova with proc glm';
     class region;
     model infrisk=region;
     means region / ;
     means region/ tukey bon scheffe;
proc plot;
     title3 'Scatterplot';
     plot infrisk * nurses
          infrisk * nurses = medschl;
proc corr;
     title3 'Correlation Matrix';
     var stay -- nbeds census nurses service;
proc glm;
     title3 'Simple regression with proc glm';
     model infrisk=nurses;


appsrv01.srv> cat basicsenic.lst
```

The FREQ Procedure

Table of region by medschl

region(Region of country (usa)) medschl(Medical school affiliation)

Frequency / Expected / Row Pct	Yes	No	Total
Northeast	6 4.3628 20.69	23 24.637 79.31	29
North Central	7 4.8142 21.88	25 27.186 78.13	32
South	2 5.4159 5.56	34 30.584 94.44	36
West	2 2.4071 12.50	14 13.593 87.50	16
Total	17	96	113

Statistics for Table of region by medschl

Statistic	DF	Value	Prob
Chi-Square	3	4.5084	0.2115
Likelihood Ratio Chi-Square	3	5.0108	0.1710
Mantel-Haenszel Chi-Square	1	2.3105	0.1285
Phi Coefficient		0.1997	
Contingency Coefficient		0.1959	
Cramer's V		0.1997	

WARNING: 38% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 113

Skipping the t-test output (see online text), ...

Study of the Effectiveness of Nosocomial Infection Control 3
Elementary tests on SENIC Data
One-way anova with proc glm
15:13 Sunday, September 5, 2004

The GLM Procedure

Class Level Information

Class	Levels	Values
region	4	North Central Northeast South West

Number of observations 113

Study of the Effectiveness of Nosocomial Infection Control 4
Elementary tests on SENIC Data
One-way anova with proc glm
15:13 Sunday, September 5, 2004

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

Study of the Effectiveness of Nosocomial Infection Control 5
Elementary tests on SENIC Data
One-way anova with proc glm
15:13 Sunday, September 5, 2004

The GLM Procedure

Level of region	N	infrisk Mean	infrisk 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

Study of the Effectiveness of Nosocomial Infection Control 6
Elementary tests on SENIC Data
One-way anova with proc glm
15:13 Sunday, September 5, 2004

The GLM Procedure

Tukey's Studentized Range (HSD) Test for infrisk

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	109
Error Mean Square	1.686283
Critical Value of Studentized Range	3.68984

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

region Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Northeast - North Central	0.5131	-0.3555	1.3818	
Northeast - West	0.5256	-0.5295	1.5808	
Northeast - South	1.0430	0.1976	1.8884	***
North Central - Northeast	-0.5131	-1.3818	0.3555	
North Central - West	0.0125	-1.0249	1.0499	
North Central - South	0.5299	-0.2933	1.3530	
West - Northeast	-0.5256	-1.5808	0.5295	
West - North Central	-0.0125	-1.0499	1.0249	
West - South	0.5174	-0.5006	1.5354	
South - Northeast	-1.0430	-1.8884	-0.1976	***
South - North Central	-0.5299	-1.3530	0.2933	
South - West	-0.5174	-1.5354	0.5006	