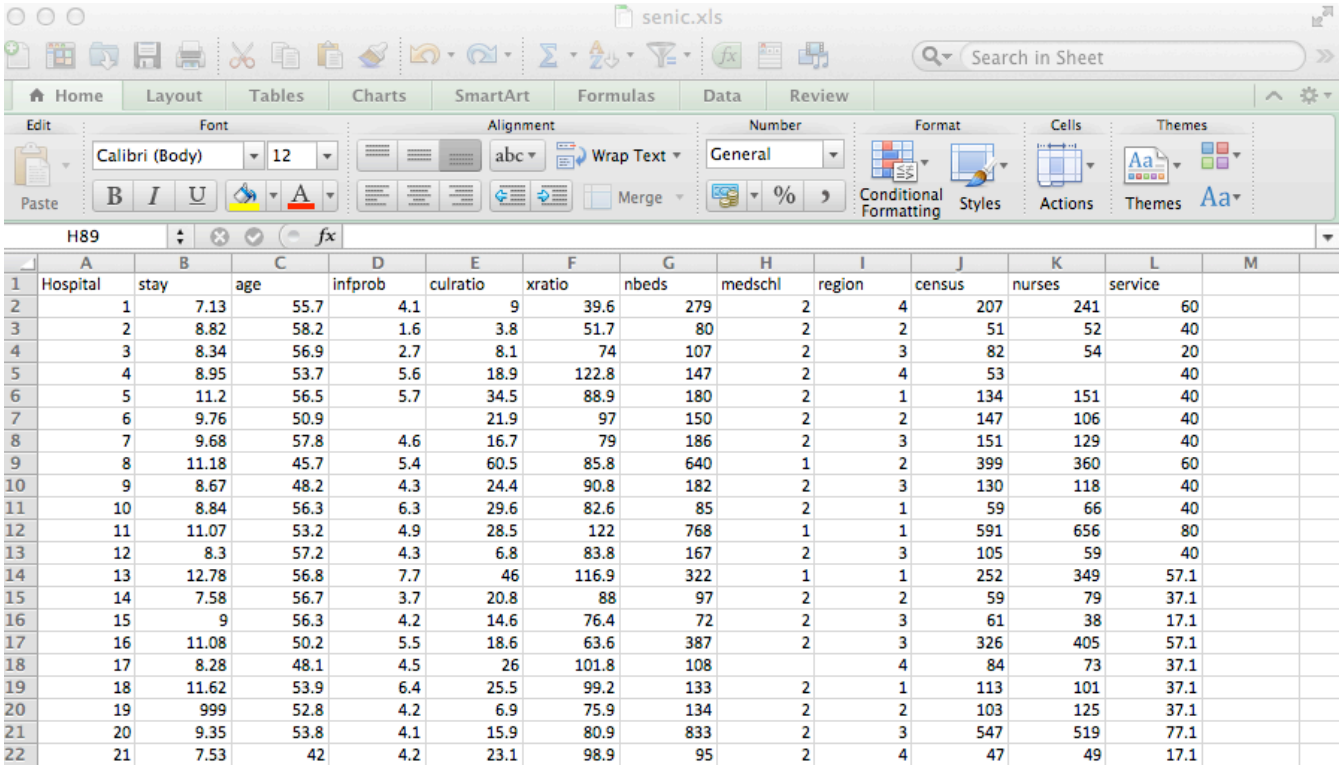
Read and Describe the SENIC Data



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Hospital	stay	age	infprob	culratio	xratio	nbeds	medschl	region	census	nurses	service	
2	1	7.13	55.7	4.1	9	39.6	279	2	4	207	241	60	
3	2	8.82	58.2	1.6	3.8	51.7	80	2	2	51	52	40	
4	3	8.34	56.9	2.7	8.1	74	107	2	3	82	54	20	
5	4	8.95	53.7	5.6	18.9	122.8	147	2	4	53		40	
6	5	11.2	56.5	5.7	34.5	88.9	180	2	1	134	151	40	
7	6	9.76	50.9		21.9	97	150	2	2	147	106	40	
8	7	9.68	57.8	4.6	16.7	79	186	2	3	151	129	40	
9	8	11.18	45.7	5.4	60.5	85.8	640	1	2	399	360	60	
10	9	8.67	48.2	4.3	24.4	90.8	182	2	3	130	118	40	
11	10	8.84	56.3	6.3	29.6	82.6	85	2	1	59	66	40	
12	11	11.07	53.2	4.9	28.5	122	768	1	1	591	656	80	
13	12	8.3	57.2	4.3	6.8	83.8	167	2	3	105	59	40	
14	13	12.78	56.8	7.7	46	116.9	322	1	1	252	349	57.1	
15	14	7.58	56.7	3.7	20.8	88	97	2	2	59	79	37.1	
16	15	9	56.3	4.2	14.6	76.4	72	2	3	61	38	17.1	
17	16	11.08	50.2	5.5	18.6	63.6	387	2	3	326	405	57.1	
18	17	8.28	48.1	4.5	26	101.8	108		4	84	73	37.1	
19	18	11.62	53.9	6.4	25.5	99.2	133	2	1	113	101	37.1	
20	19	999	52.8	4.2	6.9	75.9	134	2	2	103	125	37.1	
21	20	9.35	53.8	4.1	15.9	80.9	833	2	3	547	519	77.1	
22	21	7.53	42	4.2	23.1	98.9	95	2	4	47	49	17.1	

If the data come in an Excel spreadsheet (very common), blanks are ideal for missing values.

The spreadsheet must be .xls, not .xlsx.

Beware of trying to read a .csv file into SAS in a unix/linux environment. It may be a plain text file, but the Windows line breaks cause terrible problems.

It's *much* better for the spreadsheet to contain raw data only -- no computed variables. Do the computation with SAS.

```
% curl http://www.utstat.toronto.edu/~brunner/appliedf12/data/senic.xls > senic.xls
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 40448  100 40448    0     0  2692k      0 --:--:-- --:--:-- --:--:-- 9875k
% ls
senic.xls

% emacs senic0.sas
```

```
/* senic0.sas */
options linesize=79 pagesize=500 noovp formdlim=' ';
/* Read data from MS Excel spreadsheet */

proc import datafile="senic.xls" out=senic1 dbms=xls;
            getnames=yes;
/* Input data file is senic.xls
   Ouput data set is called senic1
   dbms=xls The input file is an Excel spreadsheet.
            Necessary to read an Excel spreadsheet directly under unix/linux
            Works under Windows too except for Excel 4.0 spreadsheets
            The xlsx file type is not supported as of SAS Version 9.2
            If there are multiple sheets, use sheet="sheet1" or something.
   getnames=yes   Use column names as variable names                    */


/* Variables are    Hospital stay age infprob culratio xratio nbeds medschl
                    region census nurses service */

proc freq;
      tables _all_;

/* Problems in SAS data set senic1
    one stay = 999
    two age=99 */


/* Once you do a proc, the data step is over. Start a new data step,
creating a new data set that can be modified. */

data senic2;
     set senic1;
/* Fix missing values */
     if stay=999 then stay = . ; /* Dot is missing for numeric */
     if age=99 then age = . ;

proc means;
     var stay age;
```

If you must have missing values that are not blank, try to make them numeric, as in this example. If there are non-numeric codes (like X or NA) for a numeric variable, the variable will be read as character. You can convert character variables to numeric if most of the “characters” are numbers, but it's a bit ugly.

This applies if you are reading data from a spreadsheet. If the data are in a plain text file, the input statement is very powerful and can handle non-numeric missing value codes for numeric data.

Parts of senic0.lst

The FREQ Procedure

Hospital

Hospital	Frequency	Percent	Cumulative Frequency	Cumulative 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
. . .				
110	1	0.88	110	97.35
111	1	0.88	111	98.23
112	1	0.88	112	99.12
113	1	0.88	113	100.00

stay

stay	Frequency	Percent	Cumulative Frequency	Cumulative Percent
6.7	1	0.88	1	0.88
7.08	1	0.88	2	1.77
7.13	1	0.88	3	2.65
7.14	3	2.65	6	5.31
7.39	1	0.88	7	6.19
7.53	1	0.88	8	7.08
7.58	1	0.88	9	7.96
7.63	1	0.88	10	8.85
7.65	1	0.88	11	9.73
7.67	1	0.88	12	10.62
7.7	1	0.88	13	11.50
7.78	1	0.88	14	12.39
. . .				
12.78	1	0.88	108	95.58
13.59	1	0.88	109	96.46
13.95	1	0.88	110	97.35
17.94	1	0.88	111	98.23
19.56	1	0.88	112	99.12
999	1	0.88	113	100.00

age

age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
38.8	1	0.88	1	0.88
42	1	0.88	2	1.77
43.7	1	0.88	3	2.65
44.2	1	0.88	4	3.54
45	1	0.88	5	4.42
45.2	1	0.88	6	5.31

. . .

62.2	1	0.88	108	95.58
63.9	1	0.88	109	96.46
64.1	1	0.88	110	97.35
65.9	1	0.88	111	98.23
99	2	1.77	113	100.00

infprob

infprob	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
2.1	1	0.88	9	7.96

. . .

6.6	1	0.91	107	97.27
7.6	1	0.91	108	98.18
7.7	1	0.91	109	99.09
7.8	1	0.91	110	100.00

Frequency Missing = 3

culratio

culratio	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1.6	1	0.88	1	0.88
1.9	1	0.88	2	1.77
2.2	1	0.88	3	2.65

. . .

42	1	0.88	110	97.35
46	1	0.88	111	98.23
52.4	1	0.88	112	99.12
60.5	1	0.88	113	100.00

The FREQ Procedure

xratio

xratio	Frequency	Percent	Cumulative Frequency	Cumulative Percent
39.6	1	0.88	1	0.88
40.4	1	0.88	2	1.77
42.6	1	0.88	3	2.65
45.7	1	0.88	4	3.54
46.5	1	0.88	5	4.42
. . .				
99.2	1	0.88	96	84.96
101	1	0.88	97	85.84
101.3	1	0.88	98	86.73
101.8	1	0.88	99	87.61
102.1	1	0.88	100	88.50
103.6	1	0.88	101	89.38
104.3	1	0.88	102	90.27
105.3	1	0.88	103	91.15
108.7	1	0.88	104	92.04
111.7	1	0.88	105	92.92
112.6	1	0.88	106	93.81
113.1	1	0.88	107	94.69
113.7	1	0.88	108	95.58
116.9	2	1.77	110	97.35
122	1	0.88	111	98.23
122.8	1	0.88	112	99.12
133.5	1	0.88	113	100.00

nbeds

nbeds	Frequency	Percent	Cumulative Frequency	Cumulative Percent
29	1	0.88	1	0.88
52	1	0.88	2	1.77
56	1	0.88	3	2.65
60	1	0.88	4	3.54
. . .				
752	1	0.88	109	96.46
768	1	0.88	110	97.35
831	1	0.88	111	98.23
833	1	0.88	112	99.12
835	1	0.88	113	100.00

medschl

medschl	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	17	15.04	17	15.04
2	94	83.19	111	98.23
X	2	1.77	113	100.00

region

region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	28	24.78	28	24.78
2	32	28.32	60	53.10
3	37	32.74	97	85.84
4	16	14.16	113	100.00

census

census	Frequency	Percent	Cumulative Frequency	Cumulative Percent
20	1	0.88	1	0.88
37	2	1.77	3	2.65
38	1	0.88	4	3.54
39	1	0.88	5	4.42
. . .				
581	1	0.88	110	97.35
591	1	0.88	111	98.23
595	1	0.88	112	99.12
791	1	0.88	113	100.00

nurses

nurses	Frequency	Percent	Cumulative Frequency	Cumulative Percent
14	1	0.89	1	0.89
19	1	0.89	2	1.79
21	1	0.89	3	2.68
22	1	0.89	4	3.57
. . .				
528	1	0.89	110	98.21
629	1	0.89	111	99.11
656	1	0.89	112	100.00

Frequency Missing = 1

service

service	Frequency	Percent	Cumulative Frequency	Cumulative Percent
5.7	1	0.88	1	0.88
11.4	1	0.88	2	1.77
14.3	1	0.88	3	2.65

. . .

77.1	1	0.88	112	99.12
80	1	0.88	113	100.00

Repeating part of the SAS program senic0.sas

```
/* Problems in SAS data set senic1
    one stay = 999
    two age=99 */

/* Once you do a proc, the data step is over. Start a new data step,
creating a new data set that can be modified. */

data senic2;
     set senic1;
/* Fix missing values */
     if stay=999 then stay = . ; /* Dot is missing for numeric */
     if age=99 then age = . ;

proc means;
     var stay age;
```

The SAS System 3

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
stay	stay	112	9.6535714	1.9192274	6.7000000	19.5600000
age	age	111	53.2441441	4.4925329	38.8000000	65.9000000

```
/************************ senicreadxls.sas  ************************
* Read the SENIC data from Excel spreadsheet, label and define some  *
* new variables. Variables in the raw data file are:                 *
* Hospital stay age infprob culratio xratio nbeds medschl region     *
* census nurses service                                              *
**********************************************************************/

title 'Study of the Effectiveness of Nosocomial Infection Control';
options linesize=79 noovp formdlim='_' ;

proc format;  /* value labels used in second data set below */
     value yesnofmt 1 = 'Yes'   0 = 'No' ;
     value regfmt  1 =  'Northeast'
                   2 =  'North Central'
                   3 =  'South'
                   4 =  'West' ;

proc import datafile="senic.xls" out=senic1 dbms=xls;
            getnames=yes;
/* Input data file is senic.xls
   Ouput data set is called senic1
   dbms=xls The input file is an Excel spreadsheet.
            Necessary to read an Excel spreadsheet directly under unix/linux
            Works under Windows too except for Excel 4.0 spreadsheets
            The xlsx file type is not supported as of SAS Version 9.2
            If there are multiple sheets, use sheet="sheet1" or something.
   getnames=yes   Use column names as variable names                  */

/* Create a new SAS data set in which data can be modified. */

data senic2;
     set senic1;
     /* Fix missing values */
     if stay=999 then stay = . ; /* Dot is missing for numeric */
     if age=99 then age = . ;

     label
           stay     = 'Av length of hospital stay, in days'
           age      = 'Average patient age'
           infprob  = '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 (1=Y, 2=N)'
           region   = 'Region of country (usa)'
           census   = 'Aver # patients in hospital per day'
           nurses   = 'Aver # nurses during study period'
           service  = '% of 35 potential facil. & services' ;

     /***** recodes, computes & ifs  *****/
     /* Age category (median split) */
        if 0<age<=53 then agecat=0;
        else if age>53 then agecat=1;
        label agecat = 'Av patient age over 53';
     /* Indicator for medical school affiliation */
        if medschl=2 then mschool=0; else mschool=medschl;
        label mschool = 'Medical school affiliation';
     /* Variance-stabilizing transformation for infprob */
        infrisk = 2 * arsin(sqrt(infprob/100));
        label infrisk = 'Infection Risk';
```

```
     /* Indicator dummy variables for region (All 4) */
        if region=. then r1=.;
           else if region=1 then r1=1;
           else r1=0;
        if region=. then r2=.;
           else if region=2 then r2=1;
           else r2=0;
        if region=. then r3=.;
           else if region=3 then r3=1;
           else r3=0;
        if region=. then r4=.;
           else if region=4 then r4=1;
           else r4=0;
        label r1 =  'Northeast'
              r2 =  'North Central'
              r3 =  'South'
              r4 =  'West' ;
    /* 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";

     /* Associating variables with their printing formats */
     format agecat mschool r1-r4  yesnofmt.;
     format region  regfmt.;
```

```
/****************** seniccheck.sas ***************/
/*            Check new vars in  SENIC Data        */
/*************************************************/

%include 'senicreadxls.sas';  /* senicreadxls.sas reads data, etc.  */
title2 'Check new vars';

proc freq;
     title3 'Check new categorical variables';
     tables age*agecat / norow nocol nopercent missing;
     tables medschl*mschool / norow nocol nopercent missing;
     tables (r1-r4)*region / norow nocol nopercent missing;

proc plot;
     plot infprob*infrisk;
```

Part of seniccheck.lst

Check new vars
Check new categorical variables

The FREQ Procedure

Table of age by agecat

age(Average patient age)
agecat(Av patient age over 53)

Frequency	.	No	Yes	Total
.	2	0	0	2
38.8	0	1	0	1
42	0	1	0	1

. . .

52.2	0	1	0	1
52.3	0	1	0	1
52.4	0	1	0	1
52.5	0	1	0	1
52.8	0	4	0	4
53	0	1	0	1
53.2	0	0	4	4
53.7	0	0	1	1
53.8	0	0	3	3
53.9	0	0	2	2

. . .

64.1	0	0	1	1
65.9	0	0	1	1
Total	2	55	56	113

Check new vars
Check new categorical variables

The FREQ Procedure

Table of medschl by mschool

medschl(Medical school affiliation (1=Y, 2=N))
mschool(Medical school affiliation)

Frequency	.	No	Yes	Total
.	2	0	0	2
1	0	0	17	17
2	0	94	0	94
Total	2	94	17	113

Table of r1 by region

r1(Northeast) region(Region of country (usa))

Frequency	Northeast	North Central	South	West	Total
No	0	32	37	16	85
Yes	28	0	0	0	28
Total	28	32	37	16	113

Table of r2 by region

r2(North Central) region(Region of country (usa))

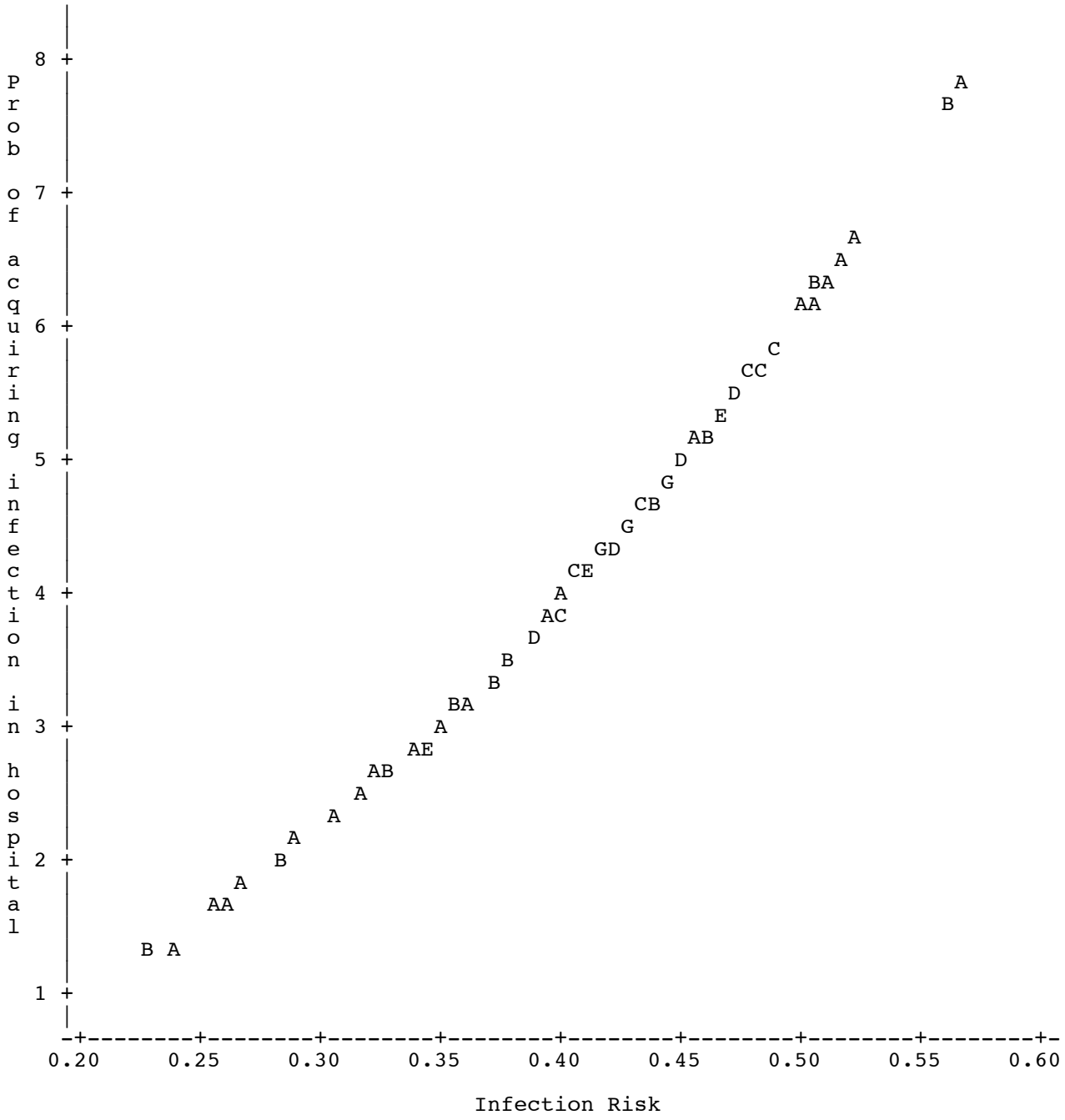
Frequency	Northeast	North Central	South	West	Total
No	28	0	37	16	81
Yes	0	32	0	0	32
Total	28	32	37	16	113

. . .

Check new vars

Check new categorical variables

Plot of infprob*infrisk. Legend: A = 1 obs, B = 2 obs, etc.



NOTE: 3 obs had missing values.

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

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

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

proc means n mean std;
     title3 'Means and SDs of quantitative variables';
     var infprob infrisk 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 infprob infrisk stay -- nbeds census nurses service;
```

Part of senicdescr.lst

```
          Study of the Effectiveness of Nosocomial Infection Control          1
                              Descriptive Statistics
                   Frequency distributions of categorical variables

                                The FREQ Procedure

                              Medical school affiliation

                                                    Cumulative    Cumulative
          mschool    Frequency     Percent     Frequency      Percent
          -----------------------------------------------------------
              No           94       84.68            94        84.68
              Yes          17       15.32           111       100.00

                              Frequency Missing = 2


                              Region of country (usa)

                                                    Cumulative    Cumulative
              region    Frequency     Percent     Frequency      Percent
          ---------------------------------------------------------------
          Northeast             28       24.78            28        24.78
          North Central         32       28.32            60        53.10
          South                 37       32.74            97        85.84
          West                  16       14.16           113       100.00


                              Av patient age over 53

                                                    Cumulative    Cumulative
           agecat    Frequency     Percent     Frequency      Percent
           -----------------------------------------------------------
              No           55       49.55            55        49.55
              Yes          56       50.45           111       100.00

                              Frequency Missing = 2
```

Study of the Effectiveness of Nosocomial Infection Control 2
Descriptive Statistics
Means and SDs of quantitative variables

The MEANS Procedure

Variable	Label	N	Mean
infprob	Prob of acquiring infection in hospital	110	4.3500000
infrisk	Infection Risk	110	0.4146363
stay	Av length of hospital stay, in days	112	9.6535714
age	Average patient age	111	53.2441441
culratio	# cultures / # no hosp acq infect	113	15.7929204
xratio	# x-rays / # no signs of pneumonia	113	81.6283186
nbeds	Average # beds during study period	113	252.1681416
census	Aver # patients in hospital per day	113	191.3716814
nurses	Aver # nurses during study period	112	173.4732143
service	% of 35 potential facil. & services	113	43.1592920

Variable	Label	Std Dev
infprob	Prob of acquiring infection in hospital	1.3571253
infrisk	Infection Risk	0.0704713
stay	Av length of hospital stay, in days	1.9192274
age	Average patient age	4.4925329
culratio	# cultures / # no hosp acq infect	10.2347074
xratio	# x-rays / # no signs of pneumonia	19.3638261
nbeds	Average # beds during study period	192.8426868
census	Aver # patients in hospital per day	153.7595639
nurses	Aver # nurses during study period	139.8705940
service	% of 35 potential facil. & services	15.2008613

Proc univariate produces lots of output. Just look at census (number of patients).

Descriptive Statistics
Describe Quantitative Variables in More Detail

The UNIVARIATE Procedure
Variable: census (Aver # patients in hospital per day)

Moments

N	113	Sum Weights	113
Mean	191.371681	Sum Observations	21625
Std Deviation	153.759564	Variance	23642.0035
Skewness	1.3793894	Kurtosis	1.73037401
Uncorrected SS	6786317	Corrected SS	2647904.39
Coeff Variation	80.346038	Std Error Mean	14.464483

Basic Statistical Measures

Location		Variability	
Mean	191.3717	Std Deviation	153.75956
Median	143.0000	Variance	23642
Mode	59.0000	Range	771.00000
		Interquartile Range	184.00000

Note: The mode displayed is the smallest of 2 modes with a count of 3.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	13.23045	Pr > |t|	<.0001
Sign	M	56.5	Pr >= |M|	<.0001
Signed Rank	S	3220.5	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.857469	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.157043	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.837031	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4.92212	Pr > A-Sq	<0.0050

Quantiles (Definition 5)

Quantile	Estimate
100% Max	791
99%	595
95%	546
90%	413
75% Q3	252
50% Median	143
25% Q1	68
10%	49

Descriptive Statistics

Describe Quantitative Variables in More Detail

The UNIVARIATE Procedure

Variable: census (Aver # patients in hospital per day)

Quantiles (Definition 5)

Quantile	Estimate
5%	40
1%	37
0% Min	20

Extreme Observations

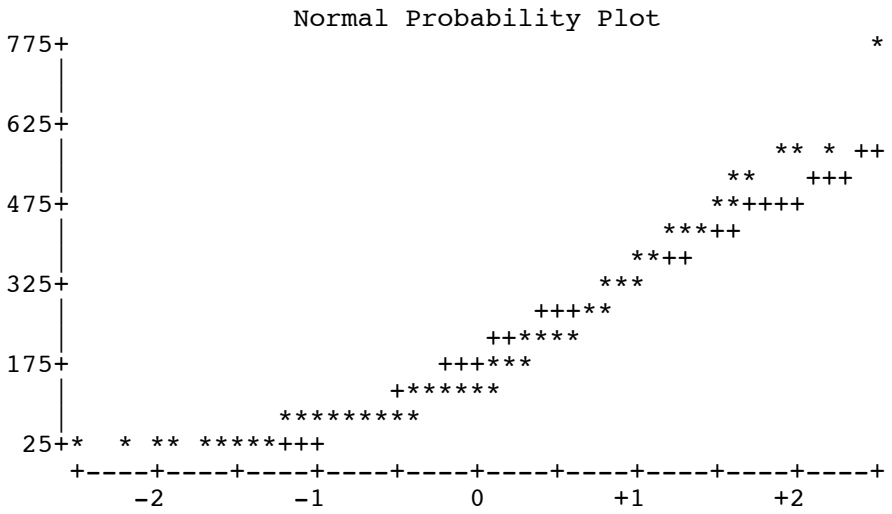
----Lowest----		----Highest---	
Value	Obs	Value	Obs
20	113	547	20
37	77	581	46
37	57	591	11
38	71	595	78
39	55	791	112

Stem	Leaf	#	Boxplot
7	9	1	0
7			
6			
6	0	1	0
5	5589	4	0
5			
4	557	3	|
4	001124	6	|
3	5679	4	|
3	11233	5	|
2	556677	6	+-----+
2	011222344444	12	| |
1	555666677889	12	| + |
1	000011112223333344444	21	*-----*
0	55555556666666667777778888999	29	+-----+
0	244444444	9	|

Multiply Stem.Leaf by 10**+2

Descriptive Statistics
Describe Quantitative Variables in More Detail

The UNIVARIATE Procedure
Variable: census (Aver # patients in hospital per day)



The list file was 36 pages long.

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

proc freq;
     title3 'Use proc freq to do crosstabs with chisquare test';
     tables mschool*region / norow nopercent 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;
 proc plot;
     title3 'Scatterplot';
     plot infrisk * nurses
          infrisk * nurses = medschl;
proc corr;
     title3 'Correlation Matrix';
     var infprob infrisk stay age census nurses nbeds service xratio
culratio;
proc reg;
     title3 'Simple regression with proc glm';
     model infrisk=quality;
```

Elementary tests on SENIC Data
Use proc freq to do crosstabs with chisquare test

The FREQ Procedure

Table of mschool by region

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

Frequency Col Pct	Northeast	North Central	South	West	Total
No	23 82.14	24 77.42	34 91.89	13 86.67	94
Yes	5 17.86	7 22.58	3 8.11	2 13.33	17
Total	28	31	37	15	111

Frequency Missing = 2

Statistics for Table of mschool by region

Statistic	DF	Value	Prob
Chi-Square	3	2.9284	0.4028
Likelihood Ratio Chi-Square	3	3.0488	0.3842
Mantel-Haenszel Chi-Square	1	1.0836	0.2979
Phi Coefficient		0.1624	
Contingency Coefficient		0.1603	
Cramer's V		0.1624	

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

Effective Sample Size = 111
Frequency Missing = 2

The TTEST Procedure

Variable: infrisk (Infection Risk)

medschl	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	17	0.4528	0.0511	0.0124	0.3423	0.5624
2	91	0.4073	0.0722	0.00757	0.2285	0.5661
Diff (1-2)		0.0455	0.0695	0.0184		

medschl	Method	Mean	95% CL Mean		Std Dev
1		0.4528	0.4266	0.4791	0.0511
2		0.4073	0.3923	0.4224	0.0722
Diff (1-2)	Pooled	0.0455	0.00909	0.0819	0.0695
Diff (1-2)	Satterthwaite	0.0455	0.0158	0.0751	

medschl	Method	95% CL Std Dev	
1		0.0380	0.0777
2		0.0631	0.0846
Diff (1-2)	Pooled	0.0612	0.0803
Diff (1-2)	Satterthwaite		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	106	2.48	0.0148
Satterthwaite	Unequal	29.478	3.13	0.0039

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	90	16	2.00	0.1170

Variable: age (Average patient age)

medschl	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	17	51.7000	3.1883	0.7733	44.2000	56.8000
2	93	53.5817	4.6390	0.4810	38.8000	65.9000
Diff (1-2)		-1.8817	4.4540	1.1748		

medschl	Method	Mean	95% CL Mean		Std Dev
1		51.7000	50.0607	53.3393	3.1883
2		53.5817	52.6263	54.5371	4.6390
Diff (1-2)	Pooled	-1.8817	-4.2105	0.4470	4.4540
Diff (1-2)	Satterthwaite	-1.8817	-3.7416	-0.0219	

medschl	Method	95% CL Std Dev	
1		2.3745	4.8523
2		4.0547	5.4217
Diff (1-2)	Pooled	3.9310	5.1388

Elementary tests on SENIC Data
T-test: Less risk at Hospitals with Med School Affiliation?

The TTEST Procedure

Variable: age (Average patient age)

medschl	Method	95% CL Std Dev
Diff (1-2)	Satterthwaite	

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	108	-1.60	0.1121
Satterthwaite	Unequal	29.999	-2.07	0.0475

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	92	16	2.12	0.0910

The GLM Procedure

Class Level Information

Class	Levels	Values
region	4	North Central Northeast South West

Number of Observations Read	113
Number of Observations Used	110

The GLM Procedure

Dependent Variable: infrisk Infection Risk

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.00979519	0.00326506	2.76	0.0460
Error	106	0.12553384	0.00118428		
Corrected Total	109	0.13532903			

R-Square	Coeff Var	Root MSE	infrisk Mean
0.072381	16.59931	0.034413	0.207318

Source	DF	Type I SS	Mean Square	F Value	Pr > F
region	3	0.00979519	0.00326506	2.76	0.0460

Source	DF	Type III SS	Mean Square	F Value	Pr > F
region	3	0.00979519	0.00326506	2.76	0.0460

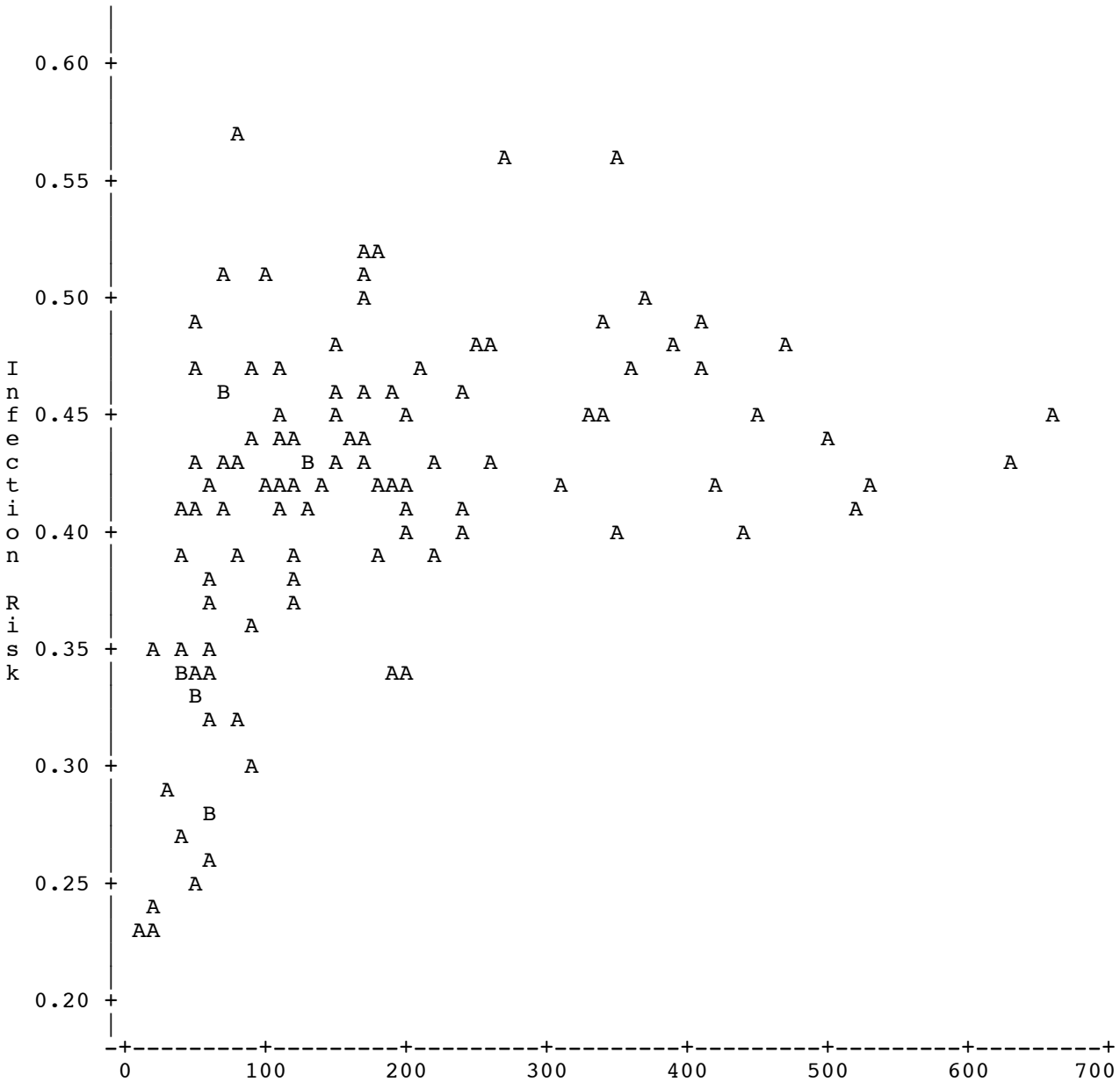
The GLM Procedure

Level of region	N	infrisk Mean	infrisk Std Dev
North Central	29	0.41573265	0.07358972
Northeast	28	0.44091219	0.06011453
South	37	0.39172997	0.07868353
West	16	0.41963742	0.04476006

Elementary tests on SENIC Data

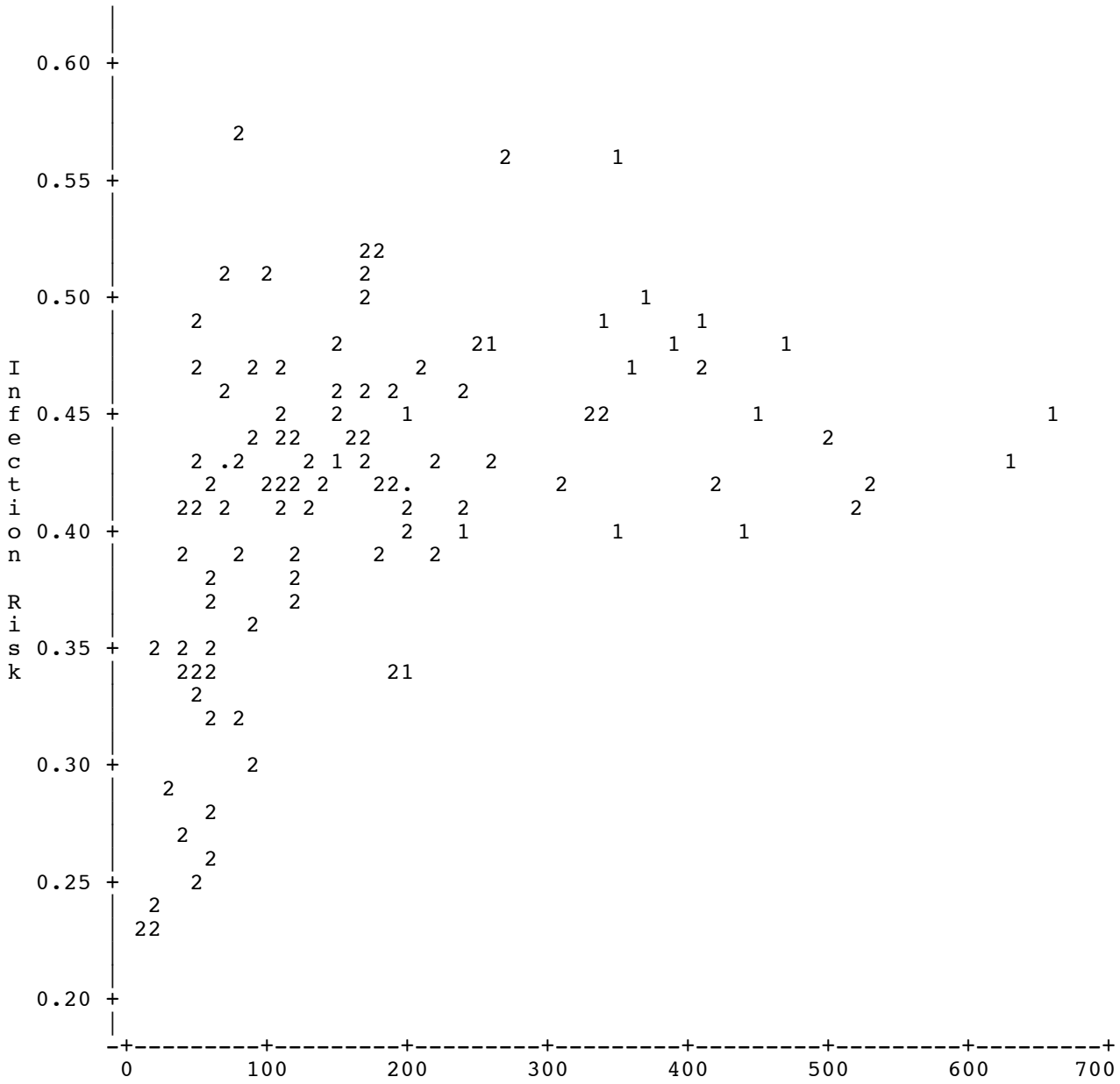
Scatterplot

Plot of infrisk*nurses. Legend: A = 1 obs, B = 2 obs, etc.



NOTE: 4 obs had missing values.

Plot of infrisk*nurses. Symbol is value of medschl.



NOTE: 4 obs had missing values. 5 obs hidden.

The CORR Procedure

10 Variables: infprob infrisk stay age census nurses nbeds service xratio culratio

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
infprob	110	4.35000	1.35713	478.50000	1.30000	7.80000
infrisk	110	0.41464	0.07047	45.61000	0.22853	0.56610
stay	112	9.65357	1.91923	1081	6.70000	19.56000
age	111	53.24414	4.49253	5910	38.80000	65.90000
census	113	191.37168	153.75956	21625	20.00000	791.00000
nurses	112	173.47321	139.87059	19429	14.00000	656.00000
nbeds	113	252.16814	192.84269	28495	29.00000	835.00000
service	113	43.15929	15.20086	4877	5.70000	80.00000
xratio	113	81.62832	19.36383	9224	39.60000	133.50000
culratio	113	15.79292	10.23471	1785	1.60000	60.50000

Simple Statistics

Variable	Label
infprob	Prob of acquiring infection in hospital
infrisk	Infection Risk
stay	Av length of hospital stay, in days
age	Average patient age
census	Aver # patients in hospital per day
nurses	Aver # nurses during study period
nbeds	Average # beds during study period
service	% of 35 potential facil. & services
xratio	# x-rays / # no signs of pneumonia
culratio	# cultures / # no hosp acq infect

The CORR Procedure

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	infprob	infrisk	stay	age
infprob Prob of acquiring infection in hospital	1.00000 110	0.99310 <.0001 110	0.53513 <.0001 109	0.00642 0.9474 108
infrisk Infection Risk	0.99310 <.0001 110	1.00000 110	0.50996 <.0001 109	-0.01878 0.8470 108
stay Av length of hospital stay, in days	0.53513 <.0001 109	0.50996 <.0001 109	1.00000 112	0.18891 0.0481 110
age Average patient age	0.00642 0.9474 108	-0.01878 0.8470 108	0.18891 0.0481 110	1.00000 111
census Aver # patients in hospital per day	0.38338 <.0001 110	0.39155 <.0001 110	0.47320 <.0001 112	-0.05833 0.5431 111
nurses Aver # nurses during study period	0.40042 <.0001 109	0.41613 <.0001 109	0.33938 0.0003 111	-0.08625 0.3703 110
nbeds Average # beds during study period	0.36308 <.0001 110	0.37256 <.0001 110	0.40843 <.0001 112	-0.06293 0.5117 111
service % of 35 potential facil. & services	0.41672 <.0001 110	0.43846 <.0001 110	0.35484 0.0001 112	-0.04726 0.6223 111
xratio # x-rays / # no signs of pneumonia	0.45168 <.0001 110	0.45675 <.0001 110	0.38198 <.0001 112	-0.01367 0.8867 111
culratio # cultures / # no hosp acq infect	0.56036 <.0001 110	0.54912 <.0001 110	0.32552 0.0005 112	-0.22489 0.0176 111

Elementary tests on SENIC Data
Correlation Matrix

The CORR Procedure

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	census	nurses	nbeds	service
infprob	0.38338	0.40042	0.36308	0.41672
Prob of acquiring infection in hospital	<.0001	<.0001	<.0001	<.0001
	110	109	110	110
infrisk	0.39155	0.41613	0.37256	0.43846
Infection Risk	<.0001	<.0001	<.0001	<.0001
	110	109	110	110
stay	0.47320	0.33938	0.40843	0.35484
Av length of hospital stay, in days	<.0001	0.0003	<.0001	0.0001
	112	111	112	112
age	-0.05833	-0.08625	-0.06293	-0.04726
Average patient age	0.5431	0.3703	0.5117	0.6223
	111	110	111	111
census	1.00000	0.90989	0.98100	0.77806
Aver # patients in hospital per day		<.0001	<.0001	<.0001
	113	112	113	113
nurses	0.90989	1.00000	0.91598	0.78343
Aver # nurses during study period	<.0001		<.0001	<.0001
	112	112	112	112
nbeds	0.98100	0.91598	1.00000	0.79452
Average # beds during study period	<.0001	<.0001		<.0001
	113	112	113	113
service	0.77806	0.78343	0.79452	1.00000
% of 35 potential facil. & services	<.0001	<.0001	<.0001	
	113	112	113	113
xratio	0.06291	0.08256	0.04582	0.11193
# x-rays / # no signs of pneumonia	0.5080	0.3868	0.6299	0.2379
	113	112	113	113
culratio	0.14295	0.19951	0.13972	0.18513
# cultures / # no hosp acq infect	0.1309	0.0349	0.1399	0.0496
	113	112	113	113

Elementary tests on SENIC Data
Correlation Matrix

The CORR Procedure

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	xratio	culratio
infprob Prob of acquiring infection in hospital	0.45168 <.0001 110	0.56036 <.0001 110
infrisk Infection Risk	0.45675 <.0001 110	0.54912 <.0001 110
stay Av length of hospital stay, in days	0.38198 <.0001 112	0.32552 0.0005 112
age Average patient age	-0.01367 0.8867 111	-0.22489 0.0176 111
census Aver # patients in hospital per day	0.06291 0.5080 113	0.14295 0.1309 113
nurses Aver # nurses during study period	0.08256 0.3868 112	0.19951 0.0349 112
nbeds Average # beds during study period	0.04582 0.6299 113	0.13972 0.1399 113
service % of 35 potential facil. & services	0.11193 0.2379 113	0.18513 0.0496 113
xratio # x-rays / # no signs of pneumonia	1.00000 113	0.42496 <.0001 113
culratio # cultures / # no hosp acq infect	0.42496 <.0001 113	1.00000 113

Elementary tests on SENIC Data
Simple regression with proc glm

The REG Procedure
Model: MODEL1
Dependent Variable: infrisk Infection Risk

Number of Observations Read	113
Number of Observations Used	106
Number of Observations with Missing Values	7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.10496	0.10496	25.26	<.0001
Error	104	0.43213	0.00416		
Corrected Total	105	0.53709			

Root MSE	0.06446	R-Square	0.1954
Dependent Mean	0.41391	Adj R-Sq	0.1877
Coeff Var	15.57357		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	0.36673	0.01128
quality	Jerry's bogus hospital quality index	1	0.00005464	0.00001087

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|
Intercept	Intercept	1	32.50	<.0001
quality	Jerry's bogus hospital quality index	1	5.03	<.0001

%