Sta442/1008f05 Overheads 2: Descriptive statistics for the math data

---

Now the math data are fairly clean. Recall that the math data have been randomly split into an exploratory sample and a relication (confirmatory) sample. The idea is to do analyses on the exploratory sample until we think we have some firm conclusions, and then try to replicate them on the confirmatory sample. The program math2.sas read the entire data set as one big file, but that was just for data cleaning. Now we will go back to just the exploratory sample. All the basic data reading and data definition will be done by `mathexread.sas`;, which is a lot like `math2.sas`, except it reads only the exploratory data. We'll just put `%include 'mathexread.sas'` near the beginning of all the files we use to do analyses. Here is a listing of `mathexread.sas`.

```
tuzo > cat mathexread.sas
/* mathexread.sas */
title 'Math Diagnostic Study:  Exploratory data';
options linesize=79 pagesize=35 noovp formdlim='_';

proc format;
     value rwfmt 0 = 'Wrong' 1 = 'Right';
     value crsfmt   4 = 'No Resp';
     value langfmt 1 = 'English'  3 = 'Other';
     value ynfmt  0 = 'No'  1 = 'Yes';
     value natfmt
        1     =   'Chinese'
        2     =   'Japanese'
        3     =   'Korean'
        4     =   'Vietnamese'
        5     =   'Other Asian'
             6     =   'Eastern European'
             7     =   'Hispanic'
             8     =   'English-speaking'
             9     =   'French'
            10     =   'Italian'
            11     =   'Greek'
            12     =   'Germanic'
            13     =   'Other European'
       14    =   'Middle-Eastern'
       15    =   'Pakistani'
       16    =   'East Indian'
       17    =   'Sub-Saharan'
       18    =   'OTHER' ;
```

```
data mathex;
     infile 'mexplore.dat';
     input id sex $ tongue nation1 nation2
           gpa english finmat alggeo hscalc
           q1-q20
           course grade;
     /* Check whether credit for HS math courses */
     if 0 <= finmat <= 100 then credfm = 1; else credfm=0;
     if 0 <= alggeo <= 100 then credag = 1; else credag=0;
     if 0 <= hscalc <= 100 then credcalc = 1; else credcalc=0;
     nhsmath = credfm+credag+credcalc;
     /* Diagnostic test subscales */
     precalc1 = sum(of q1-q4);
     precalc2 = sum(of q5-q9);
     calcone = sum(of q10-q14);
     calctwo = sum(of q15-q20);
     precalc = precalc1 + precalc2;
     calc = calcone + calctwo;
     totscore = precalc+calc;
     if english = 0 then english = .; /* Zero means mark not available */
     label
          tongue = 'Mother Tongue'
          nation1 = 'Nationality of name acc to rater1'
          nation2 = 'Nationality of name acc to rater2'
          gpa = 'High School GPA'
          english = 'Mark in HS English'
          finmat = 'Mark in HS Finite Math'
          alggeo = 'Mark in HS Algebra/Geometry'
          hscalc = 'Mark in HS Calculus'
          credfm = 'Credit for (took?) Finite math'
          credag = 'Credit for (took?) Algebra/geometry'
          credcalc = 'Credit for (took?) HS Calculus'
          nhsmath = 'Number of HS math courses'
          precalc1 = 'Precalculus 1 (bc1) subscale'
          precalc2 = 'Precalculus 2 (bc2) subscale'
          precalc  = 'Number precalculus correct'
          calcone = 'Calculus 1 (c1) subscale'
          calctwo = 'Calculus 2 (c2) subscale'
          calc  = 'Number calculus correct'
          totscore = 'Total # right on diagnostic test'
          course = 'Which university calculus course'
          grade = 'Mark in university calculus';

     /* Associate variables with printing formats */

     format q1-q20 rwfmt.;
     format course crsfmt.;
     format tongue langfmt.;
     format nation1 nation2 natfmt.;
     format credfm credag credcalc ynfmt.;

     if (50<=grade<=100) then passed=1; else passed=0;
     label passed = 'Passed the course';
     format passed ynfmt.;
```

First, basic descriptive statistics for all the variables (except identification number)

```
tuzo > cat mathdescribe1.sas
/* mathdescribe1.sas */
%include 'mathexread.sas';
title2 'Basic descriptive stats on math data';

proc freq;
     title3 'Frequency distributions of categorical variables';
     tables sex tongue nation1 nation2 q1-q20 course
            credfm credag credcalc nhsmath
            precalc1 -- totscore
            passed;

proc means n mean std;
     title3 'Means and standard deviations for quantitative variables';
     var gpa -- hscalc grade nhsmath -- totscore;

proc univariate normal plot;
     title3 'More detail for important quantitative variables';
     var gpa english hscalc totscore grade;



tuzo > sas mathdescribe1
tuzo > alias chk "grep ERROR *.log ; grep WARN *.log ; grep Invalid *.log"
tuzo > chk
tuzo >
tuzo > less mathdescribe1.lst
```

We've seen the frequency distributions of categorical variables already. Just look at the proc means output and part of the proc univariate output.

The MEANS Procedure

Variable	Label	N	Mean	Std Dev
gpa	High School GPA	466	79.2690987	6.0384921
english	Mark in HS English	480	75.8358333	8.7180498
finmat	Mark in HS Finite Math	284	78.0528169	11.1710453
alggeo	Mark in HS Algebra/Geometry	347	74.9510086	12.4299386
hscalc	Mark in HS Calculus	448	75.4441964	12.1918487
grade	Mark in university calculus	393	58.7379135	19.1831935
nhsmath	Number of HS math courses	579	1.8635579	1.0274949
precalc1	Precalculus 1 (bc1) subscale	480	2.6291667	0.8620891
precalc2	Precalculus 2 (bc2) subscale	480	1.7729167	1.2448205
calcone	Calculus 1 (c1) subscale	480	1.8375000	1.5889263
calctwo	Calculus 2 (c2) subscale	480	1.4812500	1.3319827
precalc	Number precalculus correct	480	4.4020833	1.7621620
calc	Number calculus correct	480	3.3187500	2.4918369
totscore	Total # right on diagnostic test	480	7.7208333	3.7036888

The UNIVARIATE Procedure
Variable: gpa (High School GPA)

Moments

N	466	Sum Weights	466
Mean	79.2690987	Sum Observations	36939.4
Std Deviation	6.03849212	Variance	36.4633871
Skewness	0.57602358	Kurtosis	0.06745475
Uncorrected SS	2945108.42	Corrected SS	16955.475
Coeff Variation	7.61771261	Std Error Mean	0.27972775

Basic Statistical Measures

Location		Variability	
Mean	79.26910	Std Deviation	6.03849
Median	78.30000	Variance	36.46339
Mode	76.20000	Range	32.30000
		Interquartile Range	7.30000

NOTE: The mode displayed is the smallest of 2 modes with a count of 10.

The UNIVARIATE Procedure
Variable: gpa (High School GPA)

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	283.3795	Pr > |t|	<.0001
Sign	M	233	Pr >= |M|	<.0001
Signed Rank	S	54405.5	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.972693	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.076627	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.623381	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.909041	Pr > A-Sq	<0.0050

Quantiles (Definition 5)

Quantile	Estimate
100% Max	97.3
99%	94.3
95%	91.5
90%	88.2

The UNIVARIATE Procedure
Variable: gpa (High School GPA)

Quantiles (Definition 5)

Quantile	Estimate
75% Q3	82.5
50% Median	78.3
25% Q1	75.2
10%	72.5
5%	70.7
1%	67.7
0% Min	65.0

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
65.0	44	94.3	414
66.0	380	95.0	286
66.0	2	95.8	283
67.3	366	96.2	452
67.7	446	97.3	163

Math Diagnostic Study: Exploratory data 24
Basic descriptive stats on math data
More detail for important quantitative variables
11:31 Thursday, September 1, 2005

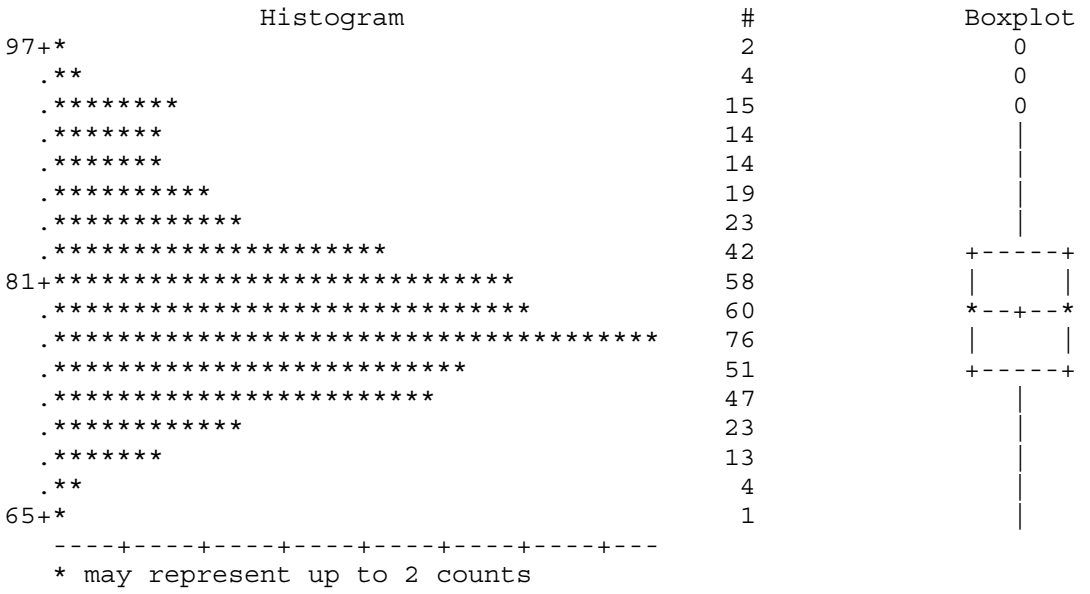
The UNIVARIATE Procedure
Variable: gpa (High School GPA)

Missing Values

Missing Value	Count	-----Percent Of----- All Obs	Missing Obs
.	113	19.52	100.00

Math Diagnostic Study: Exploratory data 25
Basic descriptive stats on math data
More detail for important quantitative variables
11:31 Thursday, September 1, 2005

The UNIVARIATE Procedure
Variable: gpa (High School GPA)



```
   Math Diagnostic Study:  Exploratory data                26
      Basic descriptive stats on math data
More detail for important quantitative variables
                                11:31 Thursday, September 1, 2005

           The UNIVARIATE Procedure
        Variable:  gpa  (High School GPA)

                 Normal Probability Plot
97+                                                      *
  |                                                    ***
  |                                                 *****+++
  |                                              ****++++
  |                                             **+++
  |                                          **++
  |                                       +**+
  |                                    +***
81+                                 *****
  |                              +****
  |                           *****
  |                      *****
  |                 *****
  |           *****+
  |   ******++
  |*** +++
65+*+++
  +----+----+----+----+----+----+----+----+----+----+
      -2         -1          0         +1         +2
```

Next, check inter-rater agreement about nationality of name.

```
tuzo > cat mathreliability.sas
```

```
/* mathreliability.sas */
%include 'mathexread.sas';
title2 'Explore inter-rater agreement about nationality of name';

proc format; /* Collapsed categories */
     value ncfmt 1 = 'Asian'
                 2 = 'European'
                 3 = 'Middle-Eastern'
                 4 = 'East Indian'
                 5 = 'Other or unknown';

options pagesize=100; /* because the nation1*nation2 table is big */

data explore2;     /* New data set */
     set mathex;   /* Copy in mathex; now they are identical. Continue... */
     if nation1=nation2 then agree1=1 ; else agree1=0;

     /* Collapse nationality categories, get better agreement */
     if           1 <= nation1 <= 5  then natcat1 = 1; /* Asian */
          else if 6 <= nation1 <= 13 then natcat1 = 2; /* European */
          else if nation1 = 14       then natcat1 = 3; /* Middle-Eastern */
          else if nation1 = 16       then natcat1 = 4; /* East Indian */
          else                            natcat1 = 5; /* Other or unknown */

     if           1 <= nation2 <= 5  then natcat2 = 1; /* Asian */
          else if 6 <= nation2 <= 13 then natcat2 = 2; /* European */
          else if nation2 = 14       then natcat2 = 3; /* Middle-Eastern */
          else if nation2 = 16       then natcat2 = 4; /* East Indian */
          else                            natcat2 = 5; /* Other or unknown */

     if natcat1=natcat2 then agree2=1 ; else agree2=0;
     format agree1 agree2 ynfmt.;
     format natcat1 natcat2 ncfmt.;

proc freq;
     title3 'Check recode of nationality'; /* Always do this */
     tables nation1*natcat1 nation2*natcat2 / norow nocol nopercent;

proc freq;
     tables agree1 agree2;

proc freq;
     tables nation1*nation2 natcat1*natcat2 / norow nocol nopercent;

/* Make a compromise variable: Call it ethnic. Rater 1 is from the middle east,
   so if he says a name is Middle-eastern, then it is. Otherwise, believe
   Rater 2 (Jerry). */

data explore3;
     set explore2;
     ethnic = natcat2;
     if natcat1 = 3 then ethnic = 3;
     label ethnic = 'Apparent ethnic background based on name';
     format ethnic ncfmt.;

proc freq;
     tables ethnic * (natcat1 natcat2) / norow nocol nopercent;
```

```
tuzo > cat mathreliability.lst
```

```
                Math Diagnostic Study:  Exploratory data                1
          Explore inter-rater agreement about nationality of name
                       Check recode of nationality
                                        11:44 Thursday, September 1, 2005

                            The FREQ Procedure

                          Table of nation1 by natcat1

    nation1(Nationality of name acc to rater1)     natcat1

    Frequency         |Asian   |European|Middle-E|East Ind|Other or|  Total
                      |        |        |astern  |ian     | unknown|
    ------------------+--------+--------+--------+--------+--------+
    Chinese           |     73 |      0 |      0 |      0 |      0 |     73
    ------------------+--------+--------+--------+--------+--------+
    Japanese          |      2 |      0 |      0 |      0 |      0 |      2
    ------------------+--------+--------+--------+--------+--------+
    Korean            |     12 |      0 |      0 |      0 |      0 |     12
    ------------------+--------+--------+--------+--------+--------+
    Vietnamese        |     16 |      0 |      0 |      0 |      0 |     16
    ------------------+--------+--------+--------+--------+--------+
    Other Asian       |     23 |      0 |      0 |      0 |      0 |     23
    ------------------+--------+--------+--------+--------+--------+
    Eastern European  |      0 |     61 |      0 |      0 |      0 |     61
    ------------------+--------+--------+--------+--------+--------+
    Hispanic          |      0 |     35 |      0 |      0 |      0 |     35
    ------------------+--------+--------+--------+--------+--------+
    English-speaking  |      0 |     96 |      0 |      0 |      0 |     96
    ------------------+--------+--------+--------+--------+--------+
    French            |      0 |      7 |      0 |      0 |      0 |      7
    ------------------+--------+--------+--------+--------+--------+
    Italian           |      0 |     22 |      0 |      0 |      0 |     22
    ------------------+--------+--------+--------+--------+--------+
    Greek             |      0 |      8 |      0 |      0 |      0 |      8
    ------------------+--------+--------+--------+--------+--------+
    Germanic          |      0 |     11 |      0 |      0 |      0 |     11
    ------------------+--------+--------+--------+--------+--------+
    Other European    |      0 |     25 |      0 |      0 |      0 |     25
    ------------------+--------+--------+--------+--------+--------+
    Middle-Eastern    |      0 |      0 |     64 |      0 |      0 |     64
    ------------------+--------+--------+--------+--------+--------+
    Pakistani         |      0 |      0 |      0 |      0 |      2 |      2
    ------------------+--------+--------+--------+--------+--------+
    East Indian       |      0 |      0 |      0 |     72 |      0 |     72
    ------------------+--------+--------+--------+--------+--------+
    Sub-Saharan       |      0 |      0 |      0 |      0 |      2 |      2
    ------------------+--------+--------+--------+--------+--------+
    OTHER             |      0 |      0 |      0 |      0 |     18 |     18
    ------------------+--------+--------+--------+--------+--------+
    Total                126      265       64       72       22      549

                          Frequency Missing = 30
```

Skipping a similar table for nationality2 ...

Math Diagnostic Study: Exploratory data 3
Explore inter-rater agreement about nationality of name
Check recode of nationality
11:44 Thursday, September 1, 2005

The FREQ Procedure

agree1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	140	24.18	140	24.18
1	439	75.82	579	100.00

agree2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	71	12.26	71	12.26
Yes	508	87.74	579	100.00

Math Diagnostic Study: Exploratory data 4
Explore inter-rater agreement about nationality of name
Check recode of nationality
11:44 Thursday, September 1, 2005

The FREQ Procedure

Table of nation1 by nation2

nation1(Nationality of name acc to rater1) × nation2(Nationality of name acc to rater2)

Frequency	Chinese	Japanese	Korean	Vietnamese	Other Asian	Eastern European	Total
Chinese	71	0	1	1	0	0	73
Japanese	0	0	0	0	0	1	2
Korean	6	0	5	0	0	0	12
Vietnamese	0	0	0	16	0	0	16
Other Asian	9	0	0	0	4	1	23
Eastern European	0	0	0	0	1	50	61
Hispanic	0	0	0	0	1	3	35
English-speaking	4	1	2	0	1	1	96

French	0	0	0	0	0	0	7
Italian	0	0	0	0	0	1	22
Greek	0	0	0	0	0	0	8
Germanic	0	0	0	0	0	1	11
Other European	0	0	0	0	0	2	25
Middle-Eastern	0	0	0	0	0	2	64
Pakistani	0	0	0	0	0	0	2
East Indian	0	0	0	0	0	1	72
Sub-Saharan	0	0	0	0	0	0	2
OTHER	6	0	0	1	1	2	18
Total	96	1	8	18	8	65	549

(Continued)

The FREQ Procedure

Table of nation1 by nation2

nation1(Nationality of name acc to rater1)
nation2(Nationality of name acc to rater2)

Frequency	Hispanic	English-speaking	French	Italian	Greek	Germanic	Total
Chinese	0	0	0	0	0	0	73
Japanese	0	0	0	0	0	0	2
Korean	0	1	0	0	0	0	12
Vietnamese	0	0	0	0	0	0	16
Other Asian	0	0	0	0	0	0	23
Eastern European	0	0	1	1	0	0	61
Hispanic	20	1	2	4	1	0	35

It goes on and on. Use scissors and tape.

Table of natcat1 by natcat2

natcat1 \ natcat2	Asian	European	Middle-Eastern	East Indian	Other or unknown	Total
Asian	113	4	2	5	2	126
European	10	245	1	4	5	265
Middle-Eastern	0	3	50	8	3	64
East Indian	0	3	1	67	1	72
Other or unknown	8	6	3	2	33	52
Total	131	261	57	86	44	579

The FREQ Procedure

Table of ethnic by natcat1

ethnic(Apparent ethnic background based on name) \ natcat1	Asian	European	Middle-Eastern	East Indian	Other or unknown	Total
Asian	113	10	0	0	8	131
European	4	245	0	3	6	258
Middle-Eastern	2	1	64	1	3	71
East Indian	5	4	0	67	2	78
Other or unknown	2	5	0	1	33	41
Total	126	265	64	72	52	579

```
                        Table of ethnic by natcat2

    ethnic(Apparent ethnic background based on name)     natcat2

    Frequency        |Asian   |European|Middle-E|East Ind|Other or|  Total
                     |        |        |astern  |ian     | unknown|
    -----------------+--------+--------+--------+--------+--------+
    Asian            |    131 |      0 |      0 |      0 |      0 |    131
    -----------------+--------+--------+--------+--------+--------+
    European         |      0 |    258 |      0 |      0 |      0 |    258
    -----------------+--------+--------+--------+--------+--------+
    Middle-Eastern   |      0 |      3 |     57 |      8 |      3 |     71
    -----------------+--------+--------+--------+--------+--------+
    East Indian      |      0 |      0 |      0 |     78 |      0 |     78
    -----------------+--------+--------+--------+--------+--------+
    Other or unknown |      0 |      0 |      0 |      0 |     41 |     41
    -----------------+--------+--------+--------+--------+--------+
    Total                131      261       57       86       44      579


tuzo >
```

It's not good for data definition to be scattered over more than one file like this. Combine `mathexread.sas` with the data step work in `mathreliability.sas`, and put it in `mathexread2.sas`. Re-run the basic descriptive stats for archival purposes using `mathdescribe2.sas`. If necessary we can keep doing this. The most current version will always have the highest number, and we can go back to an earlier version if necessary. For the record, here are `mathexread2.sas` and `mathdescribe2.sas` -- second file first to save paper.

```
/* mathdescribe2.sas */
%include 'mathexread2.sas';
title2 'Basic descriptive stats on math data';

proc freq;
     title3 'Frequency distributions of categorical variables';
      tables sex tongue nation1 nation2 agree1
                        natcat1 natcat2 agree2   ethnic
            q1-q20 course
            credfm credag credcalc nhsmath
            precalc1 -- totscore
            passed;

proc means n mean std;
     title3 'Means and standard deviations for quantitative variables';
     var gpa -- hscalc grade nhsmath -- totscore;

proc univariate normal plot;
     title3 'More detail for important quantitative variables';
     var gpa english hscalc totscore grade;
```

```
/* mathexread2.sas */
title 'Math Diagnostic Study:  Exploratory data';
options linesize=79 pagesize=35 noovp formdlim='_';

proc format;
     value rwfmt 0 = 'Wrong' 1 = 'Right';
     value crsfmt   4 = 'No Resp';
     value langfmt 1 = 'English'  3 = 'Other';
     value ynfmt  0 = 'No'  1 = 'Yes';
     value natfmt
        1     =   'Chinese'
        2     =   'Japanese'
        3     =   'Korean'
        4     =   'Vietnamese'
        5     =   'Other Asian'
             6     =   'Eastern European'
             7     =   'Hispanic'
             8     =   'English-speaking'
             9     =   'French'
            10     =   'Italian'
            11     =   'Greek'
            12     =   'Germanic'
            13     =   'Other European'
       14    =   'Middle-Eastern'
       15    =   'Pakistani'
       16    =   'East Indian'
       17    =   'Sub-Saharan'
       18    =   'OTHER' ;
     value ncfmt 1 = 'Asian'   /* Collapsed categories */
                 2 = 'European'
                 3 = 'Middle-Eastern'
                 4 = 'East Indian'
                 5 = 'Other or unknown';
data mathex;
     infile 'mexplore.dat';
     input id sex $ tongue nation1 nation2
           gpa english finmat alggeo hscalc
           q1-q20
           course grade;
     /* Check whether credit for HS math courses */
     if 0 <= finmat <= 100 then credfm = 1; else credfm=0;
     if 0 <= alggeo <= 100 then credag = 1; else credag=0;
     if 0 <= hscalc <= 100 then credcalc = 1; else credcalc=0;
     nhsmath = credfm+credag+credcalc;
     /* Diagnostic test subscales */
     precalc1 = sum(of q1-q4);
     precalc2 = sum(of q5-q9);
     calcone = sum(of q10-q14);
     calctwo = sum(of q15-q20);
     precalc = precalc1 + precalc2;
     calc = calcone + calctwo;
     totscore = precalc+calc;
     if (50<=grade<=100) then passed=1; else passed=0;
     if english = 0 then english = .; /* Zero means mark not available */

     /* Collapse nationality categories, get better agreement */
```

```
     if           1 <= nation1 <= 5  then natcat1 = 1; /* Asian */
          else if 6 <= nation1 <= 13 then natcat1 = 2; /* European */
          else if nation1 = 14       then natcat1 = 3; /* Middle-Eastern */
          else if nation1 = 16       then natcat1 = 4; /* East Indian */
          else                            natcat1 = 5; /* Other or unknown */

     if           1 <= nation2 <= 5  then natcat2 = 1; /* Asian */
          else if 6 <= nation2 <= 13 then natcat2 = 2; /* European */
          else if nation2 = 14       then natcat2 = 3; /* Middle-Eastern */
          else if nation2 = 16       then natcat2 = 4; /* East Indian */
          else                            natcat2 = 5; /* Other or unknown */
     if nation1=nation2 then agree1=1 ; else agree1=0;
     if natcat1=natcat2 then agree2=1 ; else agree2=0;

/* Make a compromise variable: Call it ethnic. Rater 1 is from the middle east,
   so if he says a name is Middle-eastern, then it is. Otherwise, believe
   Rater 2 (Jerry). */

     ethnic = natcat2;
     if natcat1 = 3 then ethnic = 3;

     label
          tongue = 'Mother Tongue'
          nation1 = 'Nationality of name acc to rater1'
          nation2 = 'Nationality of name acc to rater2'
          gpa = 'High School GPA'
          english = 'Mark in HS English'
          finmat = 'Mark in HS Finite Math'
          alggeo = 'Mark in HS Algebra/Geometry'
          hscalc = 'Mark in HS Calculus'
          credfm = 'Credit for (took?) Finite math'
          credag = 'Credit for (took?) Algebra/geometry'
          credcalc = 'Credit for (took?) HS Calculus'
          nhsmath = 'Number of HS math courses'
          precalc1 = 'Precalculus 1 (bc1) subscale'
          precalc2 = 'Precalculus 2 (bc2) subscale'
          precalc  = 'Number precalculus correct'
          calcone = 'Calculus 1 (c1) subscale'
          calctwo = 'Calculus 2 (c2) subscale'
          calc  = 'Number calculus correct'
          totscore = 'Total # right on diagnostic test'
          course = 'Which university calculus course'
          grade = 'Mark in university calculus'
          passed = 'Passed the course'
          agree1 = 'Agree on nationality of name?'
          agree2 = 'Agree on natcat?'
          ethnic = 'Apparent ethnic background based on name (compromise)';

     /* Associate variables with printing formats */
     format q1-q20 rwfmt.;
     format course crsfmt.;
     format tongue langfmt.;
     format nation1 nation2 natfmt.;
     format credfm credag credcalc passed ynfmt.;
     format agree1 agree2 ynfmt.;
     format natcat1 natcat2 ethnic ncfmt.;
     format passed ynfmt.;
```

We have been rading data in list format. The data values must be separated by one or more blanks, and all missing values must be indicated, preferably by a period. However, not all data files come like this. Suppose we had a data file in which the data are all aligned in fixed columns, missing values are indicated by blanks (this is natural; they just are not there), and some data values are packed riht up against each other. Again, this is natural if there are a lot of similar variables, like a bunch of Yes-No answers on a questionnaire, or right-wrong on a test.

The data file `mexplorefixedfmt.dat` has the same inforamtion as `mexplore.dat`. Here is the beginning of the file.

```
tuzo > less mexplorefixedfmt.dat
This is the file mexplorefixedfmt.dat: exploratory sample from the math
study. It is the same as mexplore.dat, except that it is in fixed
format, with the item scores for the diagnostic test packed into
adjacent columns.

         1111111111222222222233333333334444444445555555556666666667
12345678901234567890123456789012345678901234567890123456789012345678901234567890
 ID    SEX  T N1 N2  GPA ENGL  FM  AG  CALC     Q1-Q20         C GRADE
  1  Female 1  8  8 78.0 80.0      55  65 00010001000000000000 2 39
  2  Female 1 12 12 66.0 75.0      56  54 11111000110001000000 2 57
  3    Male 1 16 18 80.2 70.0      70  77 11100000111011000000 2 62
  4  Female 1  6  6 81.7 67.0  82  81  80 11011100001001000000 2 76
  5    Male 1 16 16 86.8 80.0  90  86  87 11010100010100000011 2 86
  6    Male 1 10 10 76.7 75.0  87      53 01011000010000000000 2 60
```

The file starts with some information that is useful, but unreadable by SAS. We would like to skip it rahter than delete it. The program `mathreadfixedfmt.sas` is just like `mathexread2.sas`, but it reads this data file. Only the infile and input statements are different. Here they are.

```
data mathex;
     infile 'mexplorefixedfmt.dat' firstobs=9 missover;
     /* firstobs=9 skips the first 8 lines in the data file. */
     /*  missover causes blanks to be missing, even at the end of a line */
     input
         #1  id       1-3    /* #1 moves pointer to beginning of record 1 */
             sex $    6-11
             tongue   13
             nation1  15-16
             nation2  18-19
             gpa      21-24
             english  26-29
             finmat   31-33
             alggeo   35-37
             hscalc   39-41  @43  /* Move pointer to column 43 */
             (q1-q20) (20*1.)     /* Read q1-q20 according to numeric format
                                     one col wide -- repeat 20 times. */
             course   64
             grade    66-67;

/* If there were a 2nd line of data, we'd continue with #2 for line 2, etc. */
```