Paper SD03

How to Use Summary Statistics as Raw Data to Do Basic Statistical Analysis

Dachao Liu, Northwestern University, Chicago, IL

ABSTRACT

Summary statistics can be N MEAN STD or even CORRELATION MATRIX. We can use them to do the same statistical analysis as we can do with the raw dataset. This will help us to save time and space if we just keep summary statistics in some situation or help us to figure out other test statistics by using summary statistics when we are reading a research paper. This paper will discuss how to use summary statistics as raw data to do basic statistical analysis.

INTRODUCTION

It's very easy to get summary statistics from a dataset, big or small, by using PROC MEANS or PROC FREQ or PROC CORR. Then we can use these summary statistics to do the same statistical analysis as we can do with the raw dataset. So we don't need to transfer the hefty size of dataset around. Just keep the summary statistics to get the job done to save time and space. This can be applied to other situations. When we are reading a research paper, sometimes we only see summary statistics. Then we can use them to get other test statistics like p-values that the raw data can produce.

DISCUSSION

Summary statistics can be N MEAN STD and CORRELATION MATRIX. For character variables, we may have N as summary statistics. N here is the number count for each category of a character variable. For numerical variables, we may have N MEAN STD or even CORRELATION MATRIX as summary statistics.

Let me start with N in character variables.

Here is a matrix of frequencies in MS Excel format. Rows are donors and columns are regions. I want to use it to do a Chi square test by reading it into SAS in the way it is in this spreadsheet.

REGION	90~100	80~90	70~80	60~70	50~60	40~50	30~40	20~30	10~20	0~10
DONOR										
h14	35	29	20	28	18	12	22	31	18	1
h25	38	25	17	10	15	14	17	15	8	4
h34	26	9	8	8	5	4	9	6	5	0
h58	44	18	18	11	17	12	27	12	3	0
h62	43	19	14	8	13	21	16	12	4	3
h66	31	11	18	8	10	14	13	7	5	0
h67	50	20	15	6	17	18	23	12	8	1
h71	34	15	17	14	16	8	14	12	3	0
h106	30	25	11	9	9	15	16	13	3	1
h111	36	22	15	19	12	22	19	14	4	0

How can I do this? First I have to shape the data to look like this:

```
1      35
1      38
1      26
1      44
1      43
1      31
1      50
1      34
1      30
1      36
2      29
2      25
2      9
2      18
2      19
2      11
2      20
2      15
```

```
2      25
2      22
.
.
.
10     1
10     4
10     0
10     0
10     3
10     0
10     1
10     0
10     1
10     0
```

I can do this data reshape in MS Excel, but it takes more effort and time. Fortunately, I can write SAS codes to do the job. First of all, I can copy the data from the Excel and paste it in a data step in SAS.

```
data onlyone;
input a b c d e f g h j k;
cards;
35     29     20     28     18     12     22     31     18     1
38     25     17     10     15     14     17     15     8      4
26     9      8      8      5      4      9      6      5      0
44     18     18     11     17     12     27     12     3      0
43     19     14     8      13     21     16     12     4      3
31     11     18     8      10     14     13     7      5      0
50     20     15     6      17     18     23     12     8      1
34     15     17     14     16     8      14     12     3      0
30     25     11     9      9      15     16     13     3      1
36     22     15     19     12     22     19     14     4      0
;
proc transpose data=onlyone out=tran;
run;

data many;
set tran;
    array rd(*)col1-col10;
    do i = 1 to dim(rd);
         count = rd(i);
         output;
    end;
keep _name_ i count;
rename _name_=region i=donor;
run;

proc format;
value donor
1='h14'
2='h25'
3='h34'
4='h58'
5='h62'
6='h66'
7='h67'
8='h71'
9='h106'
10='h111';

value $ region
      'a'='90-100'
      'b'='80-90'
      'c'='70-80'
      'd'='60-70'
      'e'='50-60'
      'f'='40-50'
      'g'='30-40'
      'h'='20-30'
      'j'='10-20'
      'k'='0-10';
```

```
run;

proc freq data=many; weight count;
table donor*region /chisq;
format donor donor. region $region.;
run;
```

The work is done. By using N statistics from character variables, I can produce other statistics like p-value that raw data can produce. In essence that is a question when I have a frequency tabulation with only N, and I want to know other Chi square statistics, like p-value.

If I have the same matrix of frequencies, I can plug my Ns and the labels in the SAS program to get Chi square test done at click of the mouse.

Now, let me discuss how to use summary statistics N MEAN STD or even CORRELATION MATRIX in numerical variables to get some statistical test done as I use the raw data.

Let's first look at TTEST.

Program 1

```
data one;
input group $ time;
cards;
a 81
a 92
a 84
a 90
a 98
b 101
b 102
b 105
b 98
b 103
;
proc ttest data=one;
 title 'data one';
 class group;
 var time;
run;
```

Program 2

```
data two;
input _stat_ $ value  group $;
cards;
  n         5         a
  mean      89        a
  std       6.7082    a
  n         5         b
  mean      101.8     b
  std       2.5884    b
  ;
proc ttest data=two;
 title 'data two';
 class group;
var value;
run;
```

I get the same output from these two programs. Please note in program 2, the data are all from the output of PROC MEANS (N MEAN STD) using data one. In other words, from some of the statistics N MEAN STD, I can produce other statistics like p-value for TTEST that raw data can produce.

Now, let's look at ANOVA.

Program 3

```
data three;
input group design;
```

```
cards;
1 18
1 20
1 17
1 15
1 16
2 35
2 33
2 39
2 27
2 32
3 25
3 25
3 23
3 26
3 23
;

proc anova;
title 'data three';
class group;
model design = group;
means group;
run;
```

Program 4;

```
data four;
input  by group   n   mean   sd;
cards;
        1   1     5  17.20  1.92353841
        1   2     5  33.20  4.38178046
        1   3     5  24.40  1.34164079
;

%inc "d:\dachao\sum_glm.sas";
       %sum_glm(data=four,
                n=n,
                mean=mean,
                stdDev=sd,
                 lsopts=stderr tdiff e,
                by=by,
                group=group)
```

I get the same output from these two programs. Please note in program 4, most of the data are from the output of PROC MEANS (N MEAN STD) using data three. In other words, from some of the statistics N MEAN STD, I can produce other statistics like analysis of variance that the raw data can produce.

There is a macro in the SAS website that has helped me to do the job. To save space, I haven't shown it in my program, instead I use %inc to include it invisibly in my program and then I use %sum_glm to call it.

This example can be extended to PROC GLM procedure when the by variable has more than two categories and we can use the same macro.

Lastly, let's look at the example of Regression.

Program 5

```
data five;
input y x1 x2 x3 x4 x5;
cards;
  1666   25   2483    472   19   448
  1696   54   2248   1339   96   694
  1063   25   3954    620   14   424
  1603   19   6565    568   38   395
  1631   39   5743   1497   36   555
  1616   45  11510   1365   25   463
  1854   65   5769   1687   45   564
  2168   56   5469   1639   47   535
  3305   96   8463   2872   88   616
```

```
  3508  127  20103   3665  182   617
  3591   93  13313   2972   64   578
  3941  134  10771   3991  106   498
  4126  129  15543   3875  128   553
;

proc reg data=five;
title 'data five';
model y = x1 x2 x3 x4 x5/;
run;
```

Program 6

```
data six(type=corr);
input  _type_ $  _name_ $   y      x1        x2         x3         x4        x5;
cards;
        mean    .      2443.692    69.76923    8610.308    2043.231    68.30769    533.8462
        std     .      1072.739    41.57755    5365.070    1276.194    49.41219    85.63765
        n       .      13.00000    13.00000    13.00000    13.00000    13.00000    13.00000
        corr    y      1.00000     0.95136     0.76076     0.96078     0.76988     0.38754
        corr    x1     0.95136     1.00000     0.77431     0.99047     0.85554     0.49150
        corr    x2     0.76076     0.77431     1.00000     0.79629     0.72921     0.20925
        corr    x3     0.96078     0.99047     0.79629     1.00000     0.82379     0.47105
        corr    x4     0.76988     0.85554     0.72921     0.82379     1.00000     0.61389
        corr    x5     0.38754     0.49150     0.20925     0.47105     0.61389     1.00000
;

proc reg data=six;
title 'data six';
model y = x1 x2 x3 x4 x5/;
run;
```

I get almost the same output from these two programs. Please note in program 6, the data are from the output of PROC CORR (N MEAN STD CORRELATION MATRIX) using data five . In other words, from some of the statistics N MEAN STD and CORRELATION MATRIX, I can produce other statistics like regression analysis that the raw data can produce.

It's very easy to get the N MEAN STD and CORRELATION MATRIX into two new data sets from PROC CORR by using two ods output statements. Then a little data manipulation can make the two data sets into data six. There is no need typing the output from PROC CORR into data six.

Besides the procedures I talked about here, there are some other SAS procedures, and I believe there will be more by the SAS Institute Inc., that can use summary statistics as an input data and produce almost the same output as the raw data can produce, like FACTOR procedure, among others.

CONCLUSION

Sometimes we have a huge dataset. If we can extract summary statistics from the dataset, we can use the summary statistics to do the same statistical analysis as we can do with the raw dataset. This extraction can be regarded as, in one sense, another kind of data reduction technique besides principal components and factor analysis. So we don't need to transfer the hefty size of dataset around. Just keep the summary statistics to get the job done to save time and space. This can be used in other situations. When we are reading a research paper, sometimes we only see summary statistics. Then we can use them to get other test statistics like p-values that the raw data can produce.

REFERENCES

1. Input Data Set of Statistics http://support.sas.com/onlinedoc/913/docMainpage.jsp
(SAS on line documentation, Base SAS, SAS/STAT: the TTEST Procedure and the REGRESSION Procedure)

2. One-way ANOVA on summary data:
http://support.sas.com/ctx/samples/index.jsp?sid=524&tab=details

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Dachao Liu
Northwestern University
Suite 1102
680 N Lake Shore Dr.
Chicago, IL 60611
Phone (312)503-2809

Email: dachao-liu@northwestern.edu