Paper RF-07-2013

Data Review Information: N-Levels or Cardinality Ratio

Ronald J. Fehd, Stakana Analytics, Atlanta, GA, USA

Abstract

Description : This paper reviews the database concept: Cardinality Ratio. The SAS(R) frequency procedure can produce an output data set with a list of the values of a variable. The number of observations of that data set is called N-Levels. The quotient of N-Levels divided by the number-of-observations of the data is the variable's Cardinality Ratio (CR). Its range is in (0–1].

Purpose : Cardinality Ratio provides an important value during data review. Four groups of values are examined.

Audience : data managers and programmers.

Programs : in this paper are available in Fehd [5, sco.Cardinality-Ratio]

Keywords : continuous, database, dimensionless, discrete, fact, frequency, keys: foreign or primary, nlevels, number of observations (nobs), unique

Quote :

Information is *the* difference
that makes *a* difference.

Gregory Bateson 1904–1980
Steps to an Ecology of Mind, 1972
italics added by R.J. Fehd

Contents

Introduction

Overview

When starting a new project a programmer or data detective may use several procedures during data discovery to confirm the relationships between variables in a data set.

The cardinality ratio (CR) of a variable is the quotient of the number of levels of that variable divided by the number of rows of the data set. The dimension: n-rows of numerator and denominator, cancels out leaving a pure number in the range of >zero – one.

CR is similar in concept to the log function: it reduces large numbers to a finite range which makes comparisons easier to grasp.

The Sets of Values

CR can be grouped into four categories.

- continuous : CR $\gtrapprox$ 0.5
- discrete: CR $\lessapprox$ 0.5
- unique: CR = 1
- single-valued: n-levels = 1

Note: One-half (0.5) is an arbitrary separation value.

continuous : information: is.a fact variable; if numeric can be summarized

discrete : indicators: character variables have standardized case: either upper or lower; numerics may be integers, or in a small finite range

information: is.a classification variable;

todo: locate one-to-one formats, or dimension (lookup) tables

unique : variable is a row-identifier; if numeric and the range is exactly 1:n-observations then it is the row-number

information: is.a primary key;

single-valued : values may be:

- character: blank
- numeric: missing
- a single value, indicating a subset

information: worthless, discard

Programs

Overview

This section examines a set of programs that produce CR.

- output to listing
- output to data set
- add CR
- analysis
- adding categories

Output: Listing

This program shows how to display the n-levels of all variables in a data set.

```
%Let lib_data  = sashelp.Class;
PROC Freq data = &lib_data
                 nlevels;
```

Output is written to the listing destination.

```
The FREQ Procedure

Number of Variable Levels

Variable      Levels
--------      ------
Name              19
Sex                2
Age                6
Height            17
Weight            15
```

Output: Data Set

Add ODS statements which create an output data set.

```
PROC Freq data     = &lib_data
                     nlevels;
          ods        exclude OneWayFreqs;
          ods        output
           nlevels = Work.Nlevels;
PROC SQL; describe table work.Nlevels;
          quit;
Proc Print data = Work.Nlevels;
run;
```

The log contains the data set description.

```
NOTE: SQL table WORK.NLEVELS was created like:
   create table WORK.NLEVELS
             (label='Number of Variable Levels')
  (TableVar char(6) label='Table Variable',
   NLevels  num     label='Number of Levels'
```

Add Cardinality Ratio This data step calculates the CR.

```
DATA Work.Card_Ratios (keep = TableVar NLevels Nobs CardRatio);
     if 0 then set &Syslast;
     attrib Nobs      length = 8
            CardRatio length = 8 label = 'Card. Ratio';
     if 0 then set &Lib_Data nobs = N_Rows;
     Nobs = N_Rows;
do until(EndoFile);
   set &Syslast end = EndoFile;
   CardRatio = Nlevels / Nobs;
   output;
   end;
stop;
run;

PROC Print label;
run;
```

Notes:

`DATA`		define output
	`if 0 set Syslast`	read data structure of Nlevels
	`attrib`	add new variables
	`if 0 set Lib-Data`	get denominator: n-rows
`do`	`until EndoFile: set`	loop: read Nlevels
	`CardRatio =`	calculate

The listing shows the Nlevels data set with the additional variables Nobs and CR.

Obs	Table Variable	Number of Levels	Nobs	Card. Ratio
1	Name	19	19	1.00000
2	Sex	2	19	0.10526
3	Age	6	19	0.31579
4	Height	17	19	0.89474
5	Weight	15	19	0.78947

The Data Here is a listing of the data set sashelp.class.

Obs	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69.0	112.5
2	Alice	F	13	56.5	84.0
3	Barbara	F	13	65.3	98.0
4	Carol	F	14	62.8	102.5
...					
16	Robert	M	12	64.8	128.0
17	Ronald	M	15	67.0	133.0
18	Thomas	M	11	57.5	85.0
19	William	M	15	66.5	112.0

Analysis Each variable may be in one of four categories:

continuous : CR $\gtrapprox$ 0.5: the variables Height and Weight have the most variation and are candidates for analysis variables

discrete : CR $\lessapprox$ 0.5: the variables Sex (gender) and Age (in years) have few values and are likely candidates for classification variables

unique : CR eq 1: by inspection of this small data set we can see that the variable Name has unique values and is the primary key.

worthless : Nlevels=1: in this data set no variables are empty

---

Adding Categories

These statements add a variable with the four category names.

```
if      Nlevels          eq 1   then is_a = 'nlevels=1';
else if CardinalityRatio eq 1   then is_a = 'primary key!?';
else if CardinalityRatio gt 0.5 then is_a = 'fact?';
else                                 is_a = 'foreign key?';
```

For data sets with many variables sorting and printing by the catagorization variable is an additional help.

```
PROC Sort  data = Work.Nlevels;
           by     is_a Name;

PROC Print data = Work.Nlevels;
           by     is_a;
           id     is_a;
```

Here is the improved listing.

```
                                          Card.
is_a             Data_Set     Name       Ratio  NLevels  Nobs

fact?          sashelp.Class  Height   0.89474       17   19
               sashelp.Class  Weight   0.78947       15   19

foreign key?   sashelp.Class  Age      0.31579        6   19
               sashelp.Class  Sex      0.10526        2   19

primary key!?  sashelp.Class  Name     1.00000       19   19
```

Summary

Conclusion

Cardinality Ratio is valuable information to have in data review. Its small finite range is easier to parse for meaning than the constantly changing and larger number-of-observations (nobs) of the data set. This difference makes a difference in understanding our data.

Further Reading

Programs : for this paper are in Fehd [5, sco.Cardinality-Ratio]

Predecessors : This paper was first published in Fehd [9, sgf2013.299].

Fehd [7, sgf2008.003] (SmryEachVar) developed a suite of programs to return a list of the frequencies of each variable in a data set or libref.

Cardinality Ratio is identified in Fehd [8, wuss2008.Database-Vocabulary] for which SmryEachVar is the predecessor.

Programs for SmryEachVar which includes calculations for Cardinality Ratio are here: Fehd [6, sco.SmryEachVar].

Theory : Contributors [4, www-wiki.dimensionless-quantity] provides the description of a ratio as a dimensionless quantity, a pure number.

Contributors [1, www-unesco.types-of-variables],
Contributors [2, www.oswego.variable-types] and
Contributors [3, www.stattrek.what-are-variables] discuss the differences between continuous and discrete variables.

References

[1] Various Contributors. Types of variables. In *Statistics*. UNESCO, 2001. URL http://www.unesco.org/webworld/idams/advguide/Chapt1_3.htm. continuous or quantitative, discrete or qualitative; online; accessed 2013-Mar-01.

[2] Various Contributors. Variable types. In *Statistics*. SUNY Oswego, 2001. URL http://www.oswego.edu/~srp/stats/variable_types.htm. categorical, explanatory ordinal, quantitative, response; online; accessed 2013-Mar-01.

[3] Various Contributors. What are variables. In *Descriptive Statistics*. Stat Trek, 2001. URL http://stattrek.com/descriptive-statistics/variables.aspx. qualitative or quantitative, continuous or discrete, univariate or bivariate; online; accessed 2013-Mar-01.

[4] Wikipedia Contributors. Dimensionless quantity. In *Wikipedia, The Free Encyclopedia*, 2013. URL http://en.wikipedia.org/w/index.php?title=Dimensionless_quantity&oldid=539704643. properties, examples and list; online; accessed 2013-Mar-01.

[5] Editor R.J. Fehd. Cardinality-ratio. In *sasCommunity.org*, 2008. URL http://www.sascommunity.org/wiki/Cardinality_Ratio. topics: definition and programs.

[6] Editor R.J. Fehd. SmryEachVar: A data-review suite for each variable in all data sets in a libref. In *sasCommunity.org*, 2008. URL http://www.sascommunity.org/wiki/SmryEachVar_A_Data_Review_Suite. list processing using parameterized includes.

[7] Ronald J. Fehd. SmryEachVar: A data-review routine for all data sets in a libref. In *Proceedings of the SAS Global Forum Annual Conference*, 2008. URL http://www2.sas.com/proceedings/forum2008/003-2008.pdf. Applications Development, 24 pp.; topics: data review; info: utilities to repair missing elements in data structure.

[8] Ronald J. Fehd. Database vocabulary: Is your data set a dimension (lookup) table, a fact table or a report? In *Western Users of SAS Software Annual Conference Proceedings*, 2008. URL http://wuss.org/proceedings08/08WUSS%20Proceedings/papers/dmw/dmw04.pdf. Databases and Warehouses, 8 pp.; topics: cardinality ratio, categories of columns (variables) and tables (data sets).

[9] Ronald J. Fehd. Data review information: N-levels or cardinality ratio. In *Proceedings of the SAS Global Forum Annual Conference*, 2013. URL http://support.sas.com/resources/papers/proceedings13/299-2013.pdf. Quick Tips, 6 pp.; topic: data review; info: theory and utilities, references.

Closure

Contact Information: **Ronald J. Fehd** mailto:Ron.Fehd.macro.maven@gmail.com
http://www.sascommunity.org/wiki/Ronald_J._Fehd

About the author:

education:	B.S. Computer Science, U/Hawaii,	1986
	SAS User Group conference attendee	since 1989
	SAS-L reader	since 1994
experience:	programmer:	25+ years
	data manager using SAS:	17+ years
	statistical software help desk:	7+ years
	author:	30+ SUG papers
	sasCommunity.org:	300+ pages
SAS-L:	author:	6,000+ messages to SAS-L since 1997
	Most Valuable SAS-L contributor:	2001, 2003

Trademarks