

A BIRDSEYE VIEW OF THE MARKET: A METHOD OF ORGANIZING, COMBINING AND DISSECTING YOUR DATA FOR MARKET ANALYSIS

C. Olivia Rud
Direct Response Group

ABSTRACT

Direct marketing has moved from a product driven, mass market approach to a market driven, custom product approach. Market segmentation commonly is used to identify subgroups which have similar geographic, demographic or lifestyle characteristics which correlate with certain performance measures such as response, risk, or profitability.

This paper presents a method of combining several data sets with similar descriptive variables but different performance measures with the goal of developing a comprehensive market overview. In this example, an internal database which houses all historical records provides accurate risk statistics. A test database on a current product configuration provides reliable response and/or paid sales rates. A census based data set provides market volumes such as number of individuals or households. Simple steps using PROC MEANS, PROC SORT, data merges and PROC PRINT generate an output from which a matrix can be developed. The matrix serves as a simple but powerful visual tool for marketers to use when developing marketing strategies.

INTRODUCTION

Business analysts are searching continually for effective ways to present findings and make recommendations to senior management. A visual presentation offers one of the best methods for conveying a set of statistical conclusions. This paper offers a step by step approach to building a market overview for supplemental health insurance. With slight modification, this technique can be applied to any market.

The choices for descriptive variables are population density, gender, income and age. The performance measurements are response rate and loss ratio (a risk measurement based on claim vs premium). Through

comparisons of different levels of response and risk combinations, profitable market sectors are identified. From these sectors a matrix is created from which managers can select specific demographic groups for product testing and solicitation.

PROCEDURE

STEP 1: Preplanning

Initial selection of the demographic variables depends on two criteria: 1) availability from all data sources and 2) predictability of your market performance. Geographic variables such as SCF (Postal Sectional Center Facility: a grouping of zipcodes) and state also can be powerfully displayed using this technique.

Performance variables can be selected from several different databases as long as the descriptive variables are available and identically scaled on each database. This example first accesses a historical database to select total premiums (totprem) and total claims (totclaim). Together, these will be used to compute a loss ratio. Next, response rates are calculated from a recent market test. Finally, population projections are estimated using a weighted sample of census data.

STEP 2: Preparation of Descriptive Variables

To control the number of distinct sets of descriptive variable combinations, consider segmenting or grouping variables into a manageable number of levels. These levels may be tailored to meet certain marketing needs or split evenly for ease of interpretation. In this example, gender (Male, Female) and population density (City, Suburban, Rural) already are coded. The remaining continuous variables are segmented through the following code:

```

data lib.history;
  set lib.history;
  if age <= 35 then triage = ' < 35 '; else
  if age <= 55 then triage = '35-55'; else
    triage = ' > 55 ';
  if income < 30000 then triincom = '$L' else
  if income < 45000 then triincom = '$M'; else
    triincom = '$H';

```

```

data lib.respond;
  set lib.respond;
  if age <= 35 then triage = ' < 35 '; else
  if age <= 55 then triage = '35-55'; else
    triage = ' > 55 ';
  if income < 30000 then triincom = '$L' else
  if income < 45000 then triincom = '$M'; else
    triincom = '$H';

```

```

data lib.census;
  set lib.census;
  if age <= 35 then triage = ' < 35 '; else
  if age <= 55 then triage = '35-55'; else
    triage = ' > 55 ';
  if income < 30000 then triincom = '$L' else
  if income < 45000 then triincom = '$M'; else
    triincom = '$H';

```

This combination of variable levels will result in $2 \times 3 \times 3 \times 3 = 54$ distinct groups.

STEP 3: Computing Group Performance Measure Means

Access to a historical database for supplemental health insurance provides average dollar amounts of claims and premiums paid for several hundred thousand customers. This database also provides demographic information for each customer. A MEANS procedure using the following code creates an output data set which contains the means of claims and premiums for each combination of population density, gender, income and age:

```

proc means data=lib.history;
  class popdens gender triincom triage;
  var totprem totclaim;
  output out=lib.histmean mean=;
  title 'Historical Data';

```

Testing on new product configurations for supplemental health insurance resulted in response rates for various demographic groups (response = 0,1). The following MEANS procedure computes the average response rate for each combination of population density, gender, income and age:

```

proc means data=lib.response;
  class popdens gender triincom triage;
  var respond;
  output out=lib.respmean mean=;
  title 'Response Data';

```

To determine market size, a MEANS procedure is performed on a weighted sample of census type data. The purpose is to generate SUM values of total records for each combination of population density, gender, income and age (totrec = 1 for each record):

```

proc means data=lib.census;
  class popdens gender triincom triage;
  var totrec;
  weight weight;
  output out=lib.censmean sum=sumrec;
  title 'Census Data';

```

NOTE: Weights are necessary for census type data to represent the entire U.S. population.

The output from the means procedure shows the number of observations in each demographic combination. If sample sizes are too small to be considered statistically significant (< 30 is a good rule of thumb) then groups can be combined or omitted.

STEP 4: Selecting the Appropriate Mean Values

The output data sets produced in the MEANS procedure will deliver all combinations of class variables. To specify only the groups defined by the 4-way combination, the following procedure uses the _TYPE_ indicator (see SAS® Procedures Guide, Version 6, Third Edition, Pg. 377 for further detail):

```

data lib.histmean;
  set lib.histmean;
  where _type_ = 15;

data lib.respmean;
  set lib.respmean;
  where _type_ = 15;

data lib.censmean;
  set lib.censmean;
  where _type_ = 15;

```

STEP 5: Sort and Merge

By merging on the specific groupings of descriptive variables, we can create a data set which contains all performance measures on each combination of population density, gender, income and age:

```

proc sort data=lib.histmean;
  by popdens gender triincom triage;

proc sort data=lib.respmean;
  by popdens gender triincom triage;

proc sort data=lib.censmean;
  by popdens gender triincom triage;

data lib.allmean;
  merge lib.histmean lib.respmean lib.censmean;
  by popdens gender triincom triage;
  if totclaim = 0 then losratio = 0; else
  losratio = totclaim/totprem;

```

STEP 6: Determine Market Dynamics

Market managers generally have guidelines which are designed to determine acceptable ranges for performance variables. In this example, our 'break-even' response rate is 0.0225 and our loss ratio tolerance is 0.30. Our profitability model allows for an increase in loss ratio of 0.10 for every 0.005 increase in response rate. (The significance of this relationship will become more clear in the final analysis.) Through the following algorithm, we will establish the market segments which create our market matrix:

```

data lib.allmean;
  set lib.allmean;
  if 0 <= losratio < .15 then lossgrp = 'L'; else
  if .15 <= losratio < .25 then lossgrp = 'L/M'; else
  if .25 <= losratio < .35 then lossgrp = 'M'; else
  if .35 <= losratio < .45 then lossgrp = 'M/H'; else
  if .45 <= losratio then lossgrp = 'H';

  if 0 <= respond < .015 then respgrp = 'L'; else
  if .015 <= respond < .020 then respgrp = 'L/M'; else
  if .020 <= respond < .025 then respgrp = 'M'; else
  if .025 <= respond < .030 then respgrp = 'M/H';
  else if .030 <= respond then respgrp = 'H';

```

NOTE: The ranges for the 'M's center on the 'break-even' values with the widths designed to reflect the profitability sensitivity mentioned above.

STEP 7: Sort and Print Groupings for Matrix

A quick PROC SORT sets up the data set for printing in the defined groups. PROC PRINT using BY variables displays the demographic makeup of each of our market segments:

```

proc sort data=lib.allmean;
  by lossgrp respgrp;

proc print label uniform data=lib.allmean;
  var popdens gender triincom triage sumrec;
  by lossgrp respgrp;
  title 'Loss Ratio vs Response Rate';

```

A section of the output shows the groups and their demographic members:

-----LOSSGRP=L/M RESPGRP=H-----					
OBS	POPDENS	GENDER	TRIINC	TRIAGE	SUMREC
33	R	M	\$H	< 35	122,491
34	S	M	\$H	> 55	119,828
35	S	M	\$M	< 35	99,025

-----LOSSGRP=M RESPGRP=L-----					
OBS	POPDENS	GENDER	TRIINC	TRIAGE	SUMREC
36	R	F	\$H	> 55	47,926
37	C	M	\$M	> 55	43,513

STEP 8: Create Matrix in Spreadsheet Software

The following page shows the final matrix which was created in Microsoft Excel®. Fine tuning in the spreadsheet software allows you to create a customized matrix which can display any number of demographic and performance variables. Shading enhances the degrees of change.

NOTE: For more efficient programming, a program using macro processing is also included.

SAS is a registered trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Author may be contacted at Direct Response Group, Phase II 4th Fl., 20 Moores Road, Frazer, PA 19380.

CONCLUSION

As you can see, the final matrix provides a considerable amount of information. The shading from 'good' to 'bad' defines the market while the double line, zig-zagging through the center, divides the profitable segments from the unprofitable segments. This takes advantage of the segment breaks. For each 0.005 increase in response rate we can tolerate a 0.10 increase in loss ratio and maintain the same profit margin. The segments that follow the diagonal from upper left to lower right have similar profitability.

Further scrutiny provides even more information. 'High Risk' segments with 'Medium' to 'Med/High' response rates represent a large number of households. To take advantage of this riskier market which is eager to do business with us, we can consider developing products with slightly higher prices or cut our costs by mailing less expensive packages.

Conversely, since 'Low Risk' groups reap higher profits by reducing claims, segments with very low risk may be persuaded to respond with a lower priced product offering or a more expensive package.

By examining clusters of certain age groups, products can be designed to target the needs of distinct markets characterized by life stage segments such as family growth or retirement years.

Any combination of demographic, geographic, and/or transaction data can be tailored to create a market matrix. An additional sample, created using SCFs (defined by the first three digits of the zipcode) shows a more geographic application of this method.

Insurance Market by State and SCF

L O S S R A T I O	Low Risk	.00 - .15	380,276		505,754		523,048		778,026		DC 200	852,647
			NJ 082, 085 VA 224 MT 595		NJ 074 VA 221, 231 CO 806 OK 734		NJ 081 MT 594 CO 802, 808, 815 AL 350, 351, 381, 364, 367, 368		VA 220, 222, 227, 229 OK 747 CO 800, 801, 812		AL 354-356, 358, 360, 363 KY 406, 408, 413 MT 590 NE 684, 687, 691	
	Low/Med Risk	.15 - .25	98,592		215,696		752,824		657,958		341,344	
			MT 597 CO 803		TN 370		RI 029 NJ 080 VA 240 NE 688 CO 813		VA 233 KY 400, 402		NJ 075, 089 KY 410, 416, 423	
	Medium Risk	.25 - .35	81,436		142,383		463,046		552,260		457,166	
			MT 598 CO 807		CO 814, 816		VA 234 KY 407 NE 689 OK 749 CO 805		NJ 087 AL 365 TN 384 NE 683 OK 730		NJ 073, 086 VA 236, 245	
Med/High Risk	.35 - .45	73,494				808,450		83,248		329,487		
		VA 228 OK 736				NJ 076 OK 731 KY 412 NE 680 CO 809		NE 685 CO 810, 811		NJ 084 VA 239 TN 374 NE 681		
High Risk	.45 +	172,723		731,950		2,030,606		1,438,259		2,783,559		
		VA 223, 226, 235, 243 MT 596, 599 OK 735, 738, 739		NJ 078 VA 242 AL 357, 362 TN 377 KY 483 OK 743, 744, 748		RI 028 NJ 077, 083 NE 693 TN 371, 376, 382, 385 KY 401, 404, 418 MT 593 OK 740, 741		VA 225 VA 230, 237 AL 352, 359, 386, 389 TN 378, 380 KY 405, 420, 421, 424 MT 592 OK 737, 746		NJ 070-072, 079 VA 232, 238, 241, 244, 246 TN 372, 379, 383 TN 373, 381 KY 409, 411 KY 414, 417, 422, 426-427 MT 591 NE 686 OK 745		
		Low Response	.000 - .015	Low/Med Response	.015 - .020	Medium Response	.020 - .025	Med/High Response	.025 - .030	High Response	.030 +	

This table is designed to compare markets of similar profitability. Area above double line represents markets with acceptable profitability. Profitability remains equal if a change in response of .005 is offset by a change in loss ratio of approximately .10 (all other factors are held constant). Market size in upper right corner of each segment is number of households estimated from 2% sample of census data. States include AL, CO, DC, KY, MT, NE, NJ, OK, RI, TN, VA. High risk segments with medium to med/high response rates represent >3.5MM households where opportunities exist for pricing/product variations.

Insurance Market by Population Density, Gender, Wealth and Age

L O S S R A T I O	Low Risk .00 - .15	380,276 R F \$H < 35 R M \$H > 55	505,754 C F \$L > 55 S F \$L 35-55	523,048 S M \$H 35-55	778,026 S M \$M 35-55 S M \$L 35-55	852,647 R M \$L > 35
	Low/Med Risk .15 - .25	99,592 S M \$L > 55	215,696 S M \$M > 55	752,824 R M \$H 35-55	657,958 S F \$M 35-55	341,344 R M \$H < 35 S M \$H > 55 S M \$M < 35
	Medium Risk .25 - .35	91,439 R F \$H > 55 C M \$H > 55	142,303 R M \$M > 55	403,046 C M \$L > 55 S F \$L > 55	552,260 S F \$M > 55	457,166 R F \$L < 35 S F \$L < 35
	Med/High Risk .35 - .45	73,494 C M \$H > 35		800,490 C M \$M < 35 C M \$M 35-55	83,248 R F \$M > 35 R F \$L > 55	329,487 C F \$L 35-55
.45 +	High Risk	1,217,231 S F \$H < 35 C M \$H > 55	731,550 R F \$H 35-55	2,030,606 C F \$H 35-55 C M \$H 35-55 S F \$M < 35 S F \$H 35-55	1,436,260 C F \$H > 55 C F \$M 35-55 C F \$L < 35 S M \$L < 35 R M \$M 35-55	2,783,559 C F \$M > 55 C M \$L < 55 C F \$M < 35 R F \$M < 35 R F \$L < 55 R M \$L < 35
	Low Response .000 - .015	Low/Med Response .015 - .020	Medium Response .020 - .025	Med/High Response .025 - .030	High Response .030 +	

This table is designed to compare markets of similar profitability. Area above double line represents markets with acceptable profitability. Profitability remains equal if a change in response of .005 is offset by a change in loss ratio of approximately .10 (all other factors are held constant). Market size in upper right corner of each segment is number of households estimated from 2% sample of census data. High risk segments with medium to med/high response rates represent >3.5MM households where opportunities exist for pricing/product variations.
 C = City S = Suburban Area R = Rural Area \$H = Income >= \$45K \$M = Income < \$45K and >= \$30K \$L = Income < \$30K

COMPLETE PROGRAM USING MACRO FACILITY

```

%macro segment;
  if age <= 35 then triage = ' < 35 '; else
  if age <= 55 then triage = '35-55'; else
    triage = ' > 55 ';
  if income < 30000 then triincom = '$L' else
  if income < 45000 then triincom = '$M'; else
    triincom = '$H';
mend;

data lib.history;
  set lib.history;
  %segment;

data lib.respond;
  set lib.respond;
  %segment;

data lib.census;
  set lib.census;
  %segment;

%macro seg_vars;
popdens gender triincom triage;
%mend;

proc means data=lib.history;
  class %seg_vars;
  var totprem totclaim;
  output out=lib.histmean mean=;
  title 'Historical Data';

proc means data=lib.response;
  class %seg_vars;
  var respond;
  output out=lib.respmean mean=;
  title 'Response Data';

proc means data=lib.census;
  class %seg_vars;
  var totrec;
  weight weight;
  output out=lib.censmean sum=sumrec;
  title 'Census Data';

data lib.histmean;
  set lib.histmean;
  where _type_ = 15;

data lib.respmean;
  set lib.respmean;
  where _type_ = 15;

data lib.censmean;
  set lib.censmean;
  where _type_ = 15;

proc sort data=lib.histmean;
  by %seg_vars;

proc sort data=lib.respmean;
  by %seg_vars;

proc sort data=lib.censmean;
  by %seg_vars;

data lib.allmean;
  merge lib.histmean lib.respmean lib.censmean;
  by %seg_vars;
  if totclaim = 0 then losratio = 0; else
  losratio = totclaim/totprem;

data lib.allmean;
  set lib.allmean;
  if 0 <= losratio < .15 then lossgrp = 'L'; else
  if .15 <= losratio < .25 then lossgrp = 'L/M'; else
  if .25 <= losratio < .35 then lossgrp = 'M'; else
  if .35 <= losratio < .45 then lossgrp = 'M/H'; else
  if .45 <= losratio then lossgrp = 'H';

  if 0 <= respond < .015 then respgrp = 'L'; else
  if .015 <= respond < .020 then respgrp = 'L/M'; else
  if .020 <= respond < .025 then respgrp = 'M'; else
  if .025 <= respond < .030 then respgrp = 'M/H';
  else if .030 <= respond then respgrp = 'H';

proc sort data=lib.allmean;
  by lossgrp respgrp;

proc print label uniform data=lib.allmean;
  var popdens gender triincom triage sumrec;
  by lossgrp respgrp;
  title 'Loss Ratio vs Response Rate';

```