Between automation and exploration: SAS graphing techniques for visualization of survey data

Chong Ho Yu, Samuel DiGangi, & Angel Jannasch-Pennell
Arizona State University, Tempe AZ 85287-0101

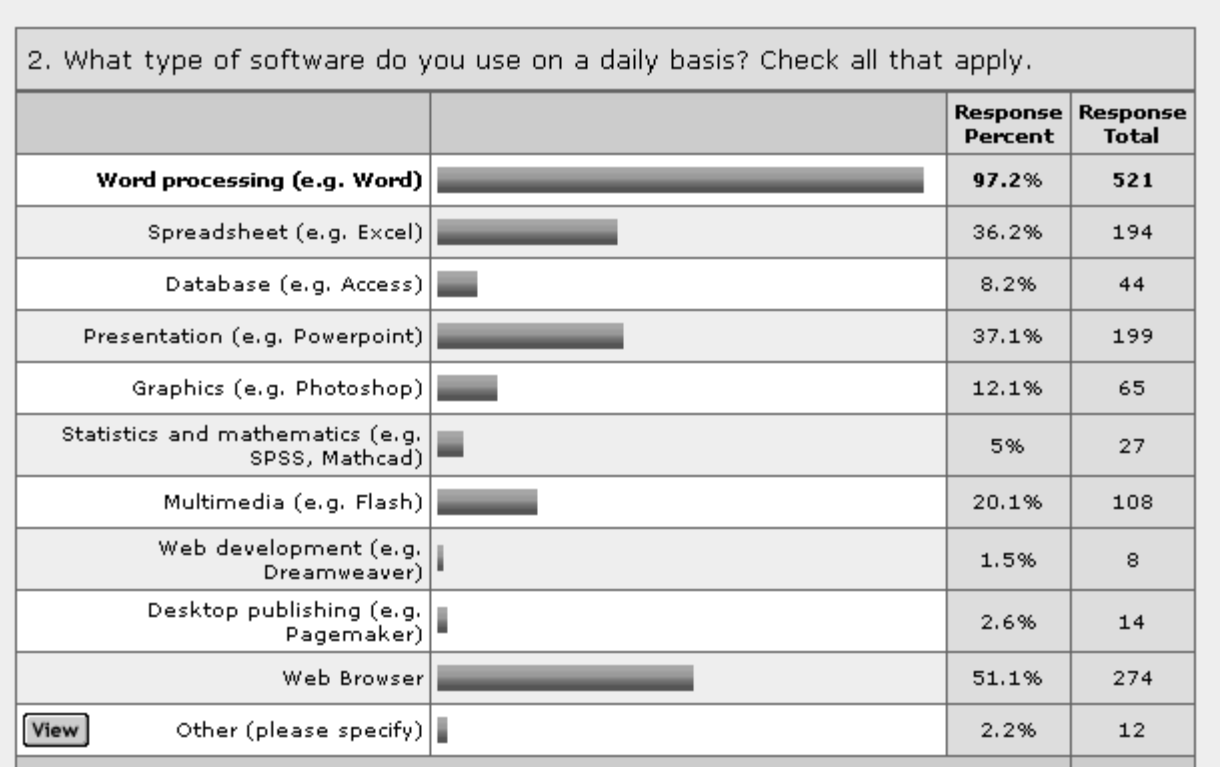
ABSTRACT

The objective of this presentation is to introduce SAS graphing techniques for visualization of survey data. There is always a tension between automation and exploration. Automation is a common practice in the context of large-scale data processing. When there are many variables, it is more efficient to automate the graphing processes via SAS/Macros and SAS/Graph. However, hidden patterns of the data may not be revealed. On the other hand, SAS/Insight, which is a tool for exploratory data analysis, requires manual manipulation of the data and thus is time-consuming. In this case study, both SAS/Macros and SAS/Insight are employed for a survey study situated Arizona State University. SAS/Macros codes are written in a way that enables the researcher to gain some degree of manipulation and exploration.

INTRODUCTION

Today, the challenge of data analysts is not the lack of tools. On the contrary, the problem is about choosing the proper tool out of numerous options. There is always a tension between automation and exploration. Automation is a common practice in the context of large-scale data processing, especially data mining. When there are many variables, it is more efficient to automate the graphing processes via SAS/Macros and SAS/Graph. However, experienced data analysts know that accepting the default settings in automated processes could result in misleading conclusions. For example, handing over our judgment to stepwise regression and some form of search algorithms has been severely criticized (Freedman, 1998). Another alarming trend is that while those automated graphs that are widely available in pre-packaged survey engines could also generate erroneous reports; most users tend to take the defaults for granted without questioning the appropriateness of those graphical presentations. This article introduces both automated and non-automated procedures as a counter-measure to the preceding problems.

Figure 1. Default histogram in a survey engine



2. What type of software do you use on a daily basis? Check all that apply.

		Response Percent	Response Total
Word processing (e.g. Word)		**97.2%**	**521**
Spreadsheet (e.g. Excel)		36.2%	194
Database (e.g. Access)		8.2%	44
Presentation (e.g. Powerpoint)		37.1%	199
Graphics (e.g. Photoshop)		12.1%	65
Statistics and mathematics (e.g. SPSS, Mathcad)		5%	27
Multimedia (e.g. Flash)		20.1%	108
Web development (e.g. Dreamweaver)		1.5%	8
Desktop publishing (e.g. Pagemaker)		2.6%	14
Web Browser		51.1%	274
View Other (please specify)		2.2%	12

INCORRECT PERCENTAGE

Recently a survey regarding usage of computing resources was conducted at Arizona State University (ASU) (DiGangi et al., 2006) using a commercial survey engine. Although the survey engine is powerful in many aspects,

the presentation does not render correct representations for questions that allow users “check all that apply.” For example, *Figure 1* shows an item as: “What type of software do you use on a daily basis? Check all that apply.” The percentages and the bars depicting the frequency counts are put side by side. It makes sense to say that 60% of the respondents are men whereas 40% are women. However, it is misleading to say X% of respondents use word processing whereas Y% use spreadsheet. This is because when subjects are allowed to check all that apply, responses are not mutually exclusive and thus the percentage will be over 100%.
As a remedy, the ASU survey team re-created all histograms using SAS/Graph and left the percentage out. For simplicity of the illustration, only 20 subjects and four options are included, as shown in Table 1.

Table 1. Responses to the question about software usage

ID	Gender	A1	A2	A3	A4
Subject 1	F	Word processing	Spreadsheet	Presentation	
Subject 2	F		Spreadsheet		
Subject 3	M	Word processing	Spreadsheet	Presentation	
Subject 4	M	Word processing			
Subject 5	M			Presentation	Other
Subject 6	M	Word processing		Presentation	
Subject 7	M		Spreadsheet		Other
Subject 8	F		Spreadsheet		
Subject 9	F	Word processing		Presentation	Other
Subject 10	F		Spreadsheet		
Subject 11	M	Word processing		Presentation	Other
Subject 12	F		Spreadsheet		
Subject 13	F				
Subject 14	M	Word processing	Spreadsheet	Presentation	
Subject 15	M	Word processing			
Subject 16	M		Spreadsheet		
Subject 17	M	Word processing	Spreadsheet	Presentation	Other
Subject 18	M	Word processing		Presentation	
Subject 19	M	Word processing		Presentation	
Subject 20	F	Word processing		Presentation	

The automated process for re-creating the histogram can be performed by three macros. An experienced SAS programmer can go even further to collapse the three steps into one. Nonetheless, it is advisable to make a logical break between tasks, because when something goes wrong, it is easier to find out where the bug is in partitioned code modules.

MACRO FOR COUNTING RESPONSES

The first macros function is "count_check." As shown below, the comments in green should be self-explanatory. It is assumed that the survey data set has been loaded into a SAS dataset named "survey."

```
/* Create a macro function to count how many people check each option

The function has two arguments: start item number, end item number */

%macro count_check(start, end);
%DO i = &start %to &end;
data temp; set survey;
      if A&i NE " " then count = 1;

/* PROC SUMMARY gives the overall count and also the count by gender */

proc summary; class gender; var count; id A&i; output out=temp&i sum = Freq;

/* To identify the option, create a new variable called "selection"
```

```
To push the option "Other" to the bottom, add a space in front of the option values */

data temp&i; set temp&i;
      if A&i NE "Other" then selection = (" "||A&i);
      else selection = A&i;
      drop A&i; run;
%END;

%mend count_check;

/* Invoke the function by starting from item A1 and end at item A4 */

%count_check(1, 4);
```

The above procedure generates four temp files, as shown in Table 2. In the first field entitled "Gender", the value for the first row is blank, because the first row is the frequency count of all subjects regardless of the gender. In PROC SUMMARY, this type of summary is called "Type 0," which is indicated in the field "_TYPE_." The field "_FREQ_" shows the sample size whereas "Freq" indicates the number of respondents who checked the option.
The next two rows, which are "Type 1 summary," show the numbers partitioned by gender.

Table 2. Temp files returned by “count_check”

Temp 1					Temp 2				
Gender	**_TYPE_**	**_FREQ_**	**Freq**	**selection**	**Gender**	**_TYPE_**	**_FREQ_**	**Freq**	**selection**
	0	20	12	Word pro…		0	20	10	Spreadsheet
F	1	8	3	Word pro…	F	1	8	5	Spreadsheet
M	1	12	9	Word pro…	M	1	12	5	Spreadsheet
Temp 3					**Temp 4**				
Gender	**_TYPE_**	**_FREQ_**	**Freq**	**selection**	**Gender**	**_TYPE_**	**_FREQ_**	**Freq**	**selection**
	0	20	11	Presentation		0	20	5	Other
F	1	8	3	Presentation	F	1	8	1	Other
M	1	12	8	Presentation	M	1	12	4	Other

MACRO FOR MERGING TEMP FILES

The second macros function is called “merge_temp”. Again, the comments embedded in the source code should be self-explanatory.

```
/* Merge separate temp files of A1 to A4 into one file */

%macro merge_temp(start, end);
data all; set
      %do i = &start %to &end;
      temp&i
      %end;
      ;run;
%mend merge_temp;

/* Invoke the function by starting from Temp1 and ending at Temp4 */

%merge_temp (1, 4);
```

The above procedure simply appends all temp files into one file. Now the data set is ready for SAS graphing, as shown in Table 3.

Table 3. Merged temp files

Gender	_TYPE_	_FREQ_	Freq	selection
	0	20	12	Word processing
F	1	8	3	Word processing
M	1	12	9	Word processing
	0	20	10	Spreadsheet
F	1	8	5	Spreadsheet
M	1	12	5	Spreadsheet
	0	20	11	Presentation
F	1	8	3	Presentation
M	1	12	8	Presentation
	0	20	5	Other
F	1	8	1	Other
M	1	12	4	Other

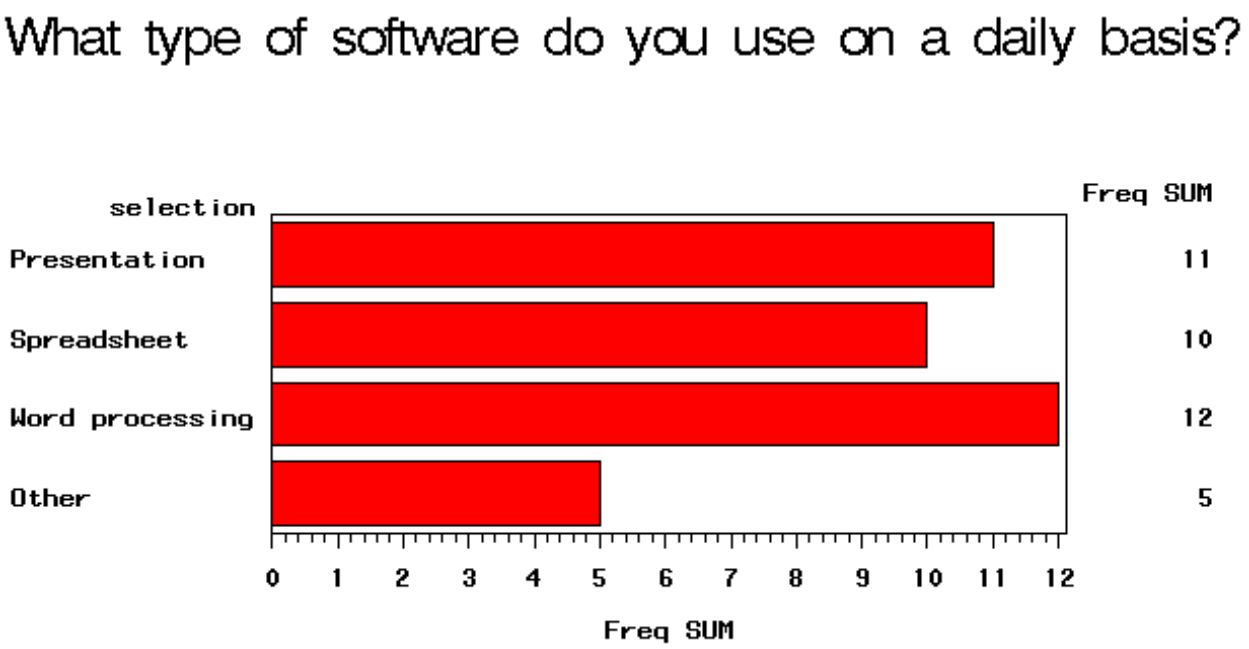
MACRO FOR PLOTTING BAR CHARTS

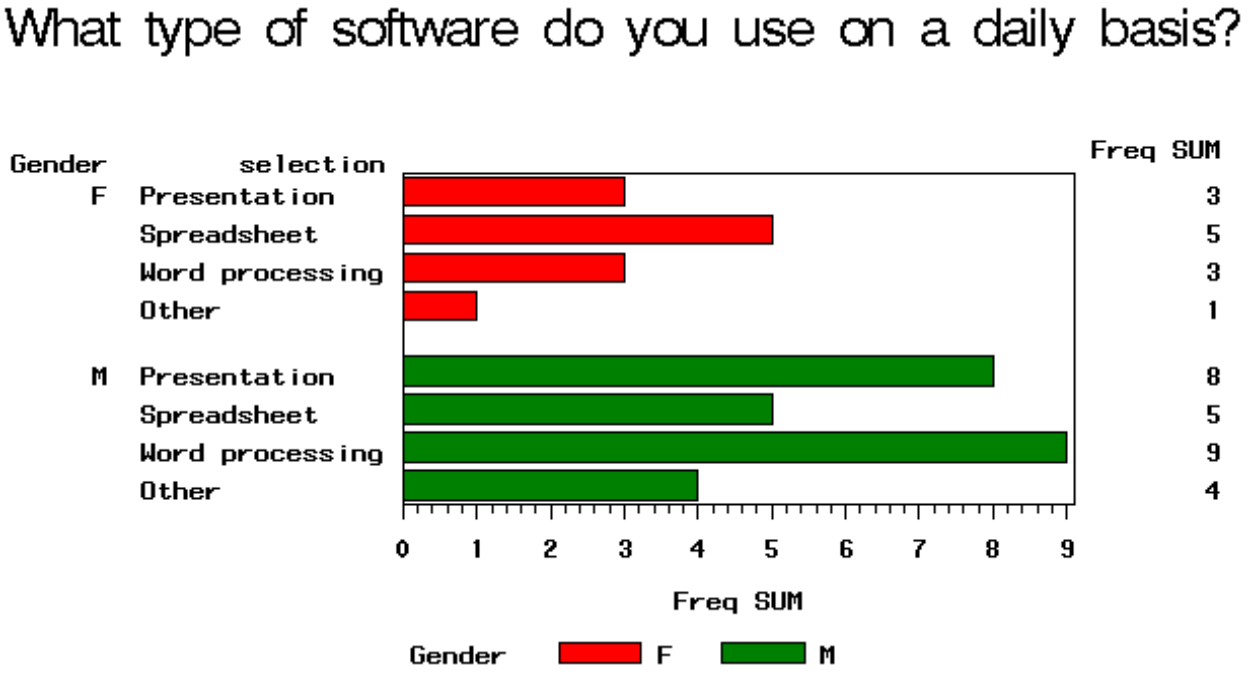
The last macros function is for graphing. This function takes four arguments:
1. Device, e.g. activex, actximg, PNG
2. File format, e.g. RTF, HTML
3. File name: Output file name, e.g. A1_A4
4. Title: Title in the graph, usually it is the text of the question.

```
/* Create a macros function to plot the graphics

%macro plot_select(device, fileformat, filename, title);

goptions reset=all device=&device;
ods &fileformat file="&filename..&fileformat" path="&path"(URL=none);

/* Plot the overall count of each option regardless of gender */

data all2; set all;
      if _TYPE_ = 0; run;
title "&title";
proc gchart data=all2;
      hbar selection /sumvar=freq;

/* Plot the frequency count of each option by gender */

data all2; set all;
      if _TYPE_ = 1; run;
proc gchart;
      hbar selection /sumvar=freq group=gender subgroup=gender; run;
ods &fileformat close; quit;
%mend plot_select;

/* Invoke the macros function, provide four arguments */

%plot_select

(PNG, html, A1_A4, Which universities did you teach in the last ten years?);
```

Figure 2 shows the graphical output without the percentage information. In this example the PNG graph format is used. To hide the “sum” next to the bar chart or to create a dynamical graph, ActiveX should be used instead of PNG. Although both PNG and GIF are static images, PNG is preferable because it usually yields a smoother image than GIF.

Figure 2. Output of SAS/macros





The beauty of SAS/Macros is alleviating repetitive tasks. To plot another item which has responses from checking all that apply, such as A5-A8, the programmer simply invoke the macros again, as shown in the following:

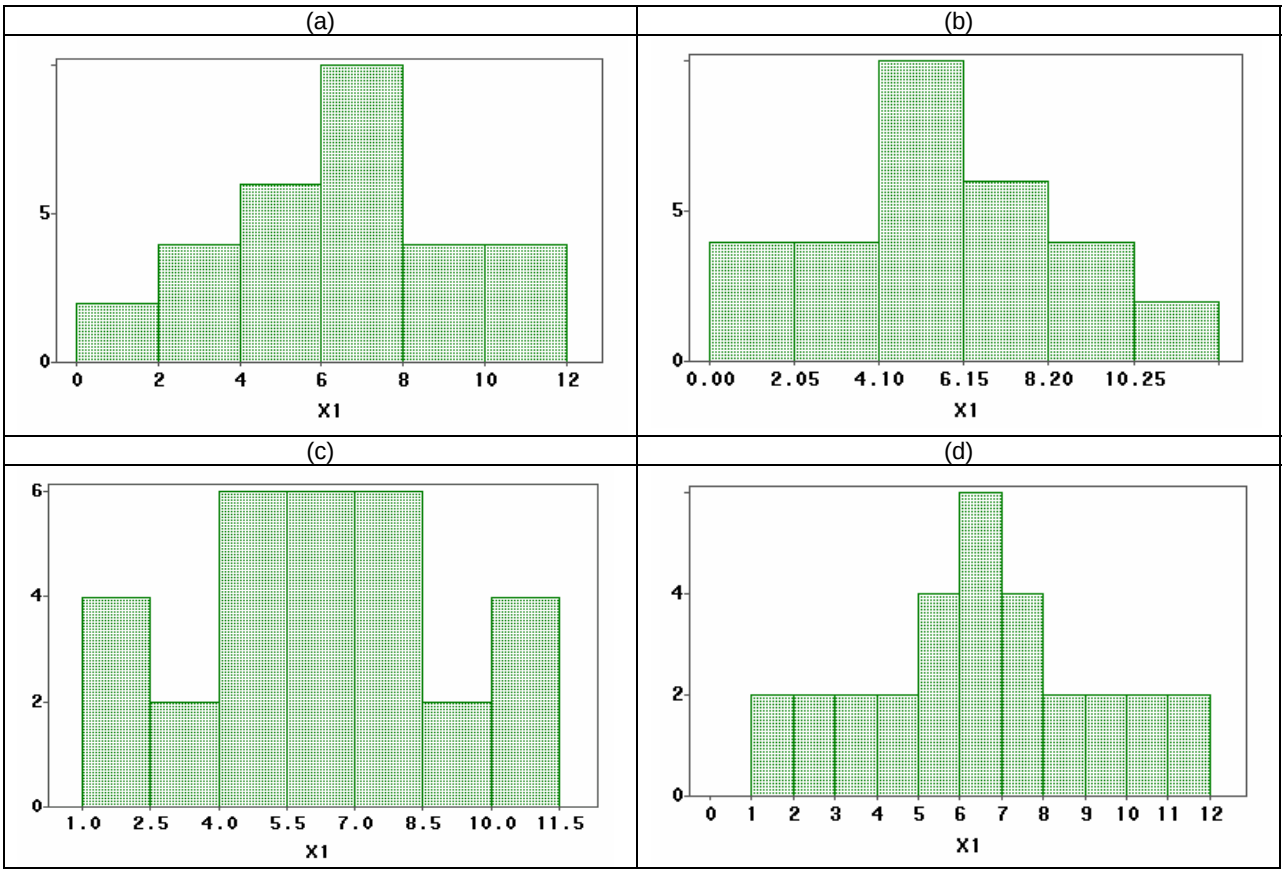
```
%count_check(5, 8);

%merge_temp (5, 8);

%plot_select (ACTXIMG, RTF, A5_A8, Which statistical packages do you use on a daily
basis?);
```

Please note that for A5-A8 the device type is changed from PNG to ActiveX Image (ACTXIMG), and the output file type is changed from HTML to RTF. The user can easily change these options without altering the original source code. Using this approach one can create hundreds of graphs to undo the misleading percentage information in a short period of time because copying, pasting, and editing are no longer necessary.

BIN-WIDTH PROBLEM

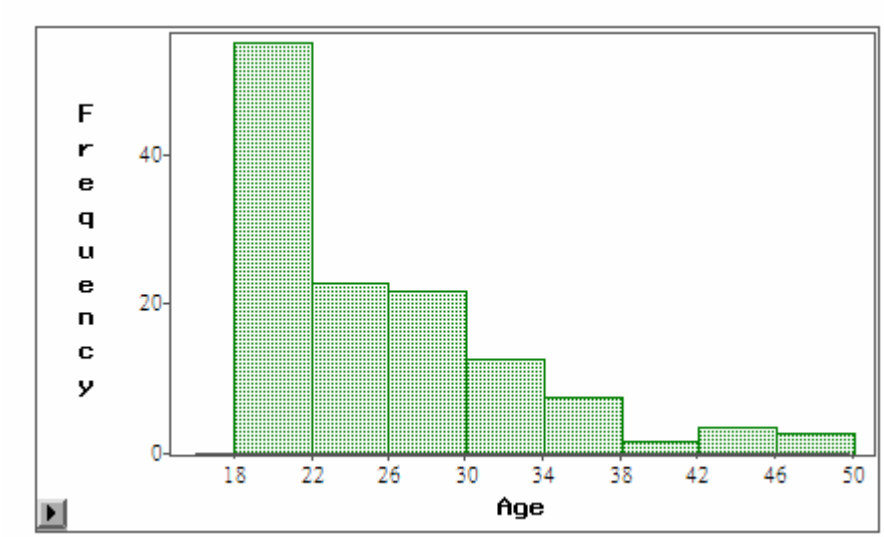
Next, we will look at another common problem found in histograms and bar charts. The appearance of a histogram or a bar chart is strongly affected by the setting of the bin-width. A bin is the "bar" in the bar chart. Bin-width refers to the size of the interval in which numbers are aggregated when determining frequencies. (The term bandwidth is similarly used in many domains including non-parametric smoothing. cf. Härdle, 1991). Different bin-widths and starting-points for bins will lead to different graphics. For example, *Figure 3*(a) is a bar chart created in SAS/Insight depicting the data vector: 1,1,2,2,3,3,4,4,5,5,5,5,6,6,6,6,6,6,7,7,7,7, 8,8,9,9,10,10,11,11. In this example, the distribution seems to skew towards the left end. However, when the bandwidth is adjusted, a totally different picture emerges: the distribution is skewed towards the right end (*Figure 3*(b)). Figure 3(c) shows a multi-modal distribution when another setting of bandwidth is used. If you are patient enough to keep adjusting the bin-width, you will eventually obtain a "normal" distribution, as shown in *Figure 4*(d) (Behrens & Yu, 2003).

Figure 3. Bar charts depicting the same data set with different bin-width settings



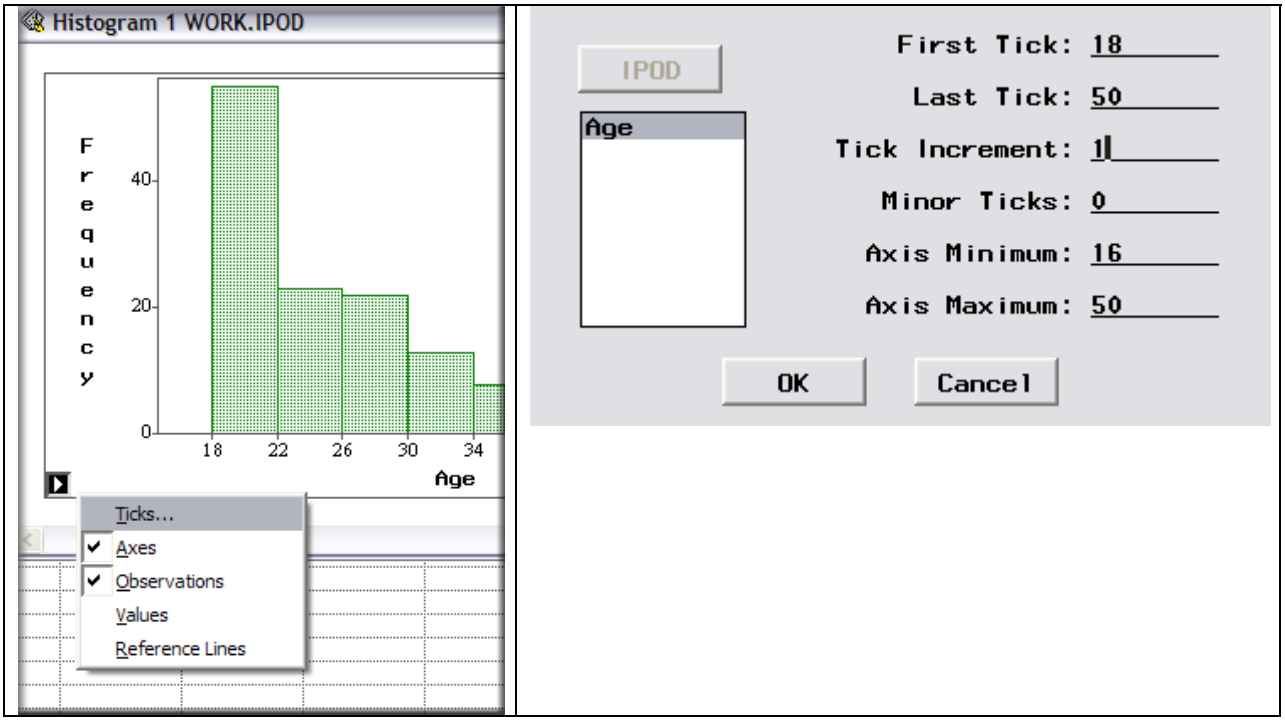
The purpose of this illustration is not to show how statistics can lie. Rather, it urges data analysts to hold a skeptical attitude toward graphical representations that are seemingly straightforward. For example, in the ASU survey there is a question pertaining to the ownership of an iPod. *Figure 4* is a bar graph generated by SAS/Insight with the default settings. It shows the number of students who own an iPod by age. At first glance, there is a clear tendency for iPod ownership to decrease as age increases.

Figure 4. Barchart of iPod ownership by age presented with the default bandwidth



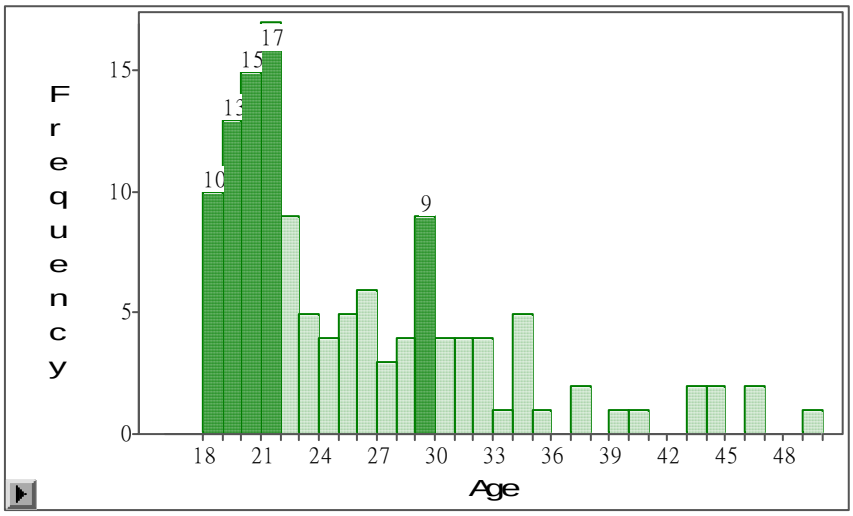
However, the graph changes dramatically after the bin-width is changed. To change the bin-width in SAS/Insight one needs to access the tick mark setting as shown in *Figure 5*. Since the "natural" breakdown of age is one year, the tick increment is changed from 4 to 1. In other words, the bin-width is 1.

Figure 5. Accessing Tick mark



After the bandwidth is altered, the data indicate that iPod ownership rises from age 18 to 22; however, after this age, ownership significantly drops (see *Figure 6*). But near age 30 the number goes up again and then goes down after 30. The possible explanation is that students who fall within the age group 18-22 financially depend on their parents and thus they can afford owning iPods. But after age 22, most students become self-reliant and hence they have to wait until age 30 for this to be financially feasible. However, in their late adult life students may lose interest in fashions like MP3 music. Henceforth, ownership depresses again. If the generation of bar charts had been automated in SAS/Macros, this insight would never have been unveiled.

Figure 6. Barchart of iPod ownership by age presented with the altered bandwidth



CONCLUSION

To automate or not to automate, that is the question! As many experienced researchers point out, data analysis is an art rather than a science. In this article, both approaches are demonstrated, but there is no clear cut criterion of when to use which approach. Nonetheless, use of both approaches in the ASU survey is motivated by the attempt to solve a common problem: the possibility of mis-representation of data by accepting default settings. While the first type of problem (sum of percents is over 100) is obvious and thus can be amended by automation, the latter (trend across age groups) is hidden and therefore requires non-automated manipulation of graphical settings.

REFERENCES

Behrens, J. T., & Yu, C. H. (2003). Exploratory data analysis. In J. A. Schinka & W. F. Velicer, (Eds.). *Handbook of psychology Volume 2: Research methods in Psychology* (pp. 33-64). New Jersey: John Wiley & Sons, Inc.

DiGangi, S., Jannasch-Pennell, A., Yu, C. H., and Kilic, Z. (2006). *1:1 Computing: Toward an understanding of technological means, needs and preferences of the ASU Downtown Phoenix Campus student*. Applied Learning Technologies Institute, Arizona State University, Tempe, AZ. Retrieved April 4, 2006 from http://www.sannier.net/wiki1/images/1/12/StudentTechSurvey-DPC.doc

Freedman, D. (1998). Rejoiner to Spirites and Scheines, In Vaughn R. McKim & Stephen P. Turner, (Eds.). *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences* (pp. 177-182), Norte Dame, IN: University of Norte Dame Press.

Härdle, W. (1991). *Smoothing techniques with implementation in S*. New York: Springer-Verlag.

CONTACT INFORMATION

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

Chong Ho Yu, Ph.D.
Director of Testing, Measurement, Assessment, and Research
Applied Learning Technology Institute
Arizona State University
3S89 Computing Commons
Tempe, AZ 85287-0101
USA
Work Phone: 480-727-0670
Email: chonghoyu@gmail.com
Web: http://www.creative-wisdom.com