Data Visualization: "Every Picture Tells a Story" or "Putting Atomic Numbers on Your Socks" ?

AnnMaria De Mars, The Julia Group, Santa Monica, CA
Revati Kadu, University of Southern California

ABSTRACT

Good data visualization is all about telling a story. Visual representation in one context is often misapplied to another, for example, maps of movies, a concept which one classic text author called as logical as putting atomic numbers on your socks. What is a bad visualization for one question can be good for another. An example we use is a map of likely voters. To forecast an election, this is misleading, because large states like Wyoming have fewer electoral votes than tiny New Jersey. However, if your question is which regions supported the Democrats, a map is a good choice. Dashboards can be a useful visualization for monitoring performance in a corporation. Scientific research requires a different set of graphics. In this paper, several examples of visual representation are used for the purpose of answering questions commonly posed in management and science. The first set of examples crosses both disciplines, using graphics to discuss data quality and outcomes in a new product evaluation. The second example uses 2008 election poll data to map Obama and McCain supporters by geography, region, electoral vote, race and income. The first two examples use graphics generated by SAS® Enterprise Guide and customized with SAS code. The third example imports a SAS dataset into JMP (just for fun) using ROC curves, decision trees and other data mining options to predict and explain post-secondary education choices.

INTRODUCTION

If you're a statistician, when you look at a table for a repeated measures Analysis of Variance, your eyes may automatically gravitate toward the F-statistic, r-square and then to the p-value for the time by treatment interaction? The fact is, though, most readers' eyes glaze over. Conveying statistical results to the general public is critically important. Statistics is not a field where we "labor each alone" . We don't do our research so we can put it on a shelf, sit in a rocking chair and feel happy with ourselves. The purpose of data analysis is to produce information that we can then share with other people in a manner that they will understand.

Steele and Iliinsky, the authors of Beautiful Visualization, make three great points:

1. Data visualization is all around us. The periodic table is a type of data visualization. So is a map of the London subway system.
2. Visual representation in one context is often misapplied to another. People have done maps of movies, technology and more. For example, this map of the 250 best movies of all time:
   http://blog.vodkaster.com/2009/06/25/the-top-250-best-movies-of-all-time-map/
   I think the authors are a bit harsh in saying that this makes about as much sense as putting atomic numbers on your socks. It is kind of an original twist if you are looking at visualization as some kind ART. However, it fails in terms of visualization of data if you think, like I do, in terms of adding information.

3. Data visualizations is really about story-telling, that is
   Question + Data + Visual = Story

Good data visualization is all about telling a story. Visual representation in one context is often misapplied to another, for example, maps of movies, a concept which one classic text author called as logical as putting atomic numbers on your socks. What is a bad visualization for one question can be good for another. Take, for example, a map of likely voters. To forecast an election, this is misleading, because large states like Wyoming have fewer electoral votes than tiny New Jersey. However, if your question is which regions supported the Democrats, a map is a good choice. Dashboards can be a useful visualization for monitoring performance in a corporation. Scientific research requires a different set of graphics.

DATA VISUALIZATION BY EXAMPLE
Example #1 WHAT WORKS? DATA VISUALIZATION IN EVALUATION RESEARCH

In government-funded programs, from education to social services to substance abuse prevention, emphasis is being placed on "What Works". Spending money on a problem is no longer enough, programs are required to document outcomes.

Much of my consulting work for the past twenty years has involved evaluations of federally funded grant programs. In this task, I come in as an outside evaluator to answer two basic questions:

1. Was the program implemented as promised? Answering this question is the goal of process evaluation. Now, I'm flying in from thousands of miles away, if unethical staff members just enter some names and test scores in their database, how will I possibly know? How can I tell in a few days if they have been accurately reporting there data.

Below are two scatter plots of data from a project that used a combined on-line and workshop format to provide staff training. There was an experimental group, which received the training, and a control group that did not. The first thing I do is to plot the pre-test against the post-test, by group. (At the end of this paper the few clicks I used to get this with SAS Enterprise Guide are given.)

If this is a reliable test, there should be a high correlation between pre- and post-test for the control group, fitting pretty close to a straight line.

If the training was effective at all, there should be a correlation between the pretest and post-test for the experimental group, but there should be more scatter around the line.

Why? Because some people benefit more from training than others. Some come late, leave early and fall asleep in between. Others pay rapt attention and read more about the topic on-line when they get home. Some people with really high scores may have known all of the information in the training and not gained a point. Other people with average pre-test scores may have learned a lot and moved up to a higher score. People who had a very high pre-test score should still have a high post-test score. Hopefully, your training didn't make them dumber

So, this is the pattern I am looking for if the program was implemented as it should have been – more scatter on the experimental group, tighter in the control group and those with high scores are more likely to stay high than low or moderate scores are to stay in the same place.

2. Did the program work? Answering this question is the goal of outcome evaluation.

There is a reference line on each plot which is the mean at the pretest, before training. In this case, using a graph might even be a little better than a repeated measures ANOVA because we have a few outliers in the experimental group. One person was nearly four standard deviations from the mean on the pretest. Another person was five standard deviations above the (pretest) mean on the post-test. Could the significant difference in favor of the experimental group be due to these outliers? The answer, as you can clearly see, is no.

You can see that the control group is about equally above and below the mean at post-test while on the right almost all of the experimental group is above the mean.

```
PROC SORT DATA=SASUSER.COPT4SPSS(KEEP=z_total_pre z_total_post group)
      OUT=wussexample;
      BY group;
GOPTIONS HBY = 2 ;
TITLE1 "Pretest by Post-Test";
FOOTNOTE1 "Funding provided by USDA Small Business Innovation Research
Award";
PROC GPLOT DATA=wussexample      UNIFORM;
PLOT z_total_post * z_total_pre /
      VREF=0 ;
      BY group;
```

Three programming points to note above –

GOPTIONS HBY = 2 ;
I wanted to highlight the difference between the control and experimental group. The default font size has the by-group title very small relative to the first title line.
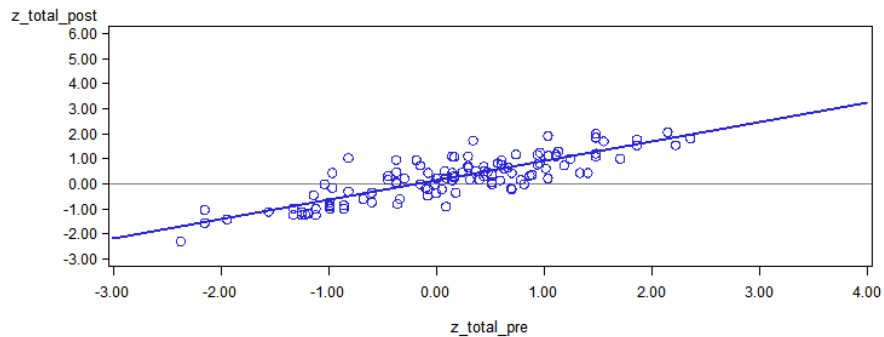
PROC GPLOT DATA = datasetname UNIFORM ;
The UNIFORM option overrides the default which is to use the values to set the minimum and maximum of the axes. To facilitate comparison, you want the experimental and control groups to have the same axes.

VREF = 0 ;
Draws a vertical line at the specified point, in this case 0. This reference line allows one to compare, at a glance, the proportions above and below the post-test mean.
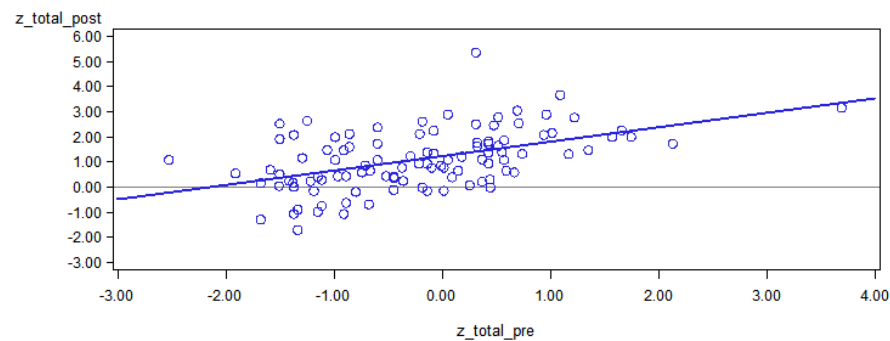
Pretest by Post-Test
group=CONTROL

FIGURE 1

Pretest by Post-Test
group=EXPERIMENTAL

FIGURE 2

Interaction Plot for z_total
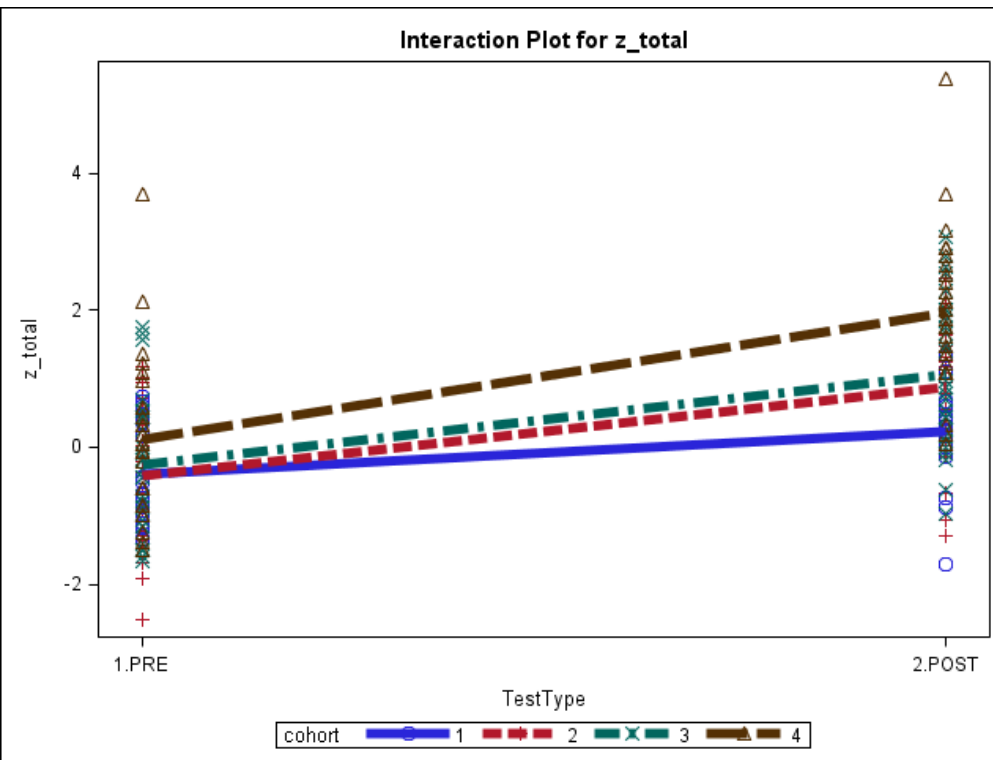
As statisticians, we never feel quite secure without equations, so look in the notes in your SAS log.

```
NOTE: Regression equation :  z_total_post =  0.13379 + 0.776552*z_total_pre.
NOTE: The above message was for the following BY group:  group=CONTROL
NOTE: Regression equation :  z_total_post =  1.233616 + 0.578418*z_total_pre.
NOTE: The above message was for the following BY group:     group=EXPERIMENTAL
```

A related question to whether the program worked is, under what conditions did it work. Statisticians immediately recognize this as a question of whether there is an interaction effect. In testing a new program such as this, it is possible that a novelty effect may exist, where the initial cohort shows much more improvement than later training. We had seen this happen in the past with computer-aided instruction in middle schools, when once the novelty of using the computer wore off, the advantage of this instructional method disappeared. Alternatively (and this is what we expected having been involved with the project), it may be that as the trainers became more experienced using the new program, as the on-site coordinators worked out logistics, the program would be more effective. In industrial engineering, it would be what we refer to as a learning curve.

The interaction plot above shows the TestType by Cohort Interaction for the Experimental Group. We can see that the first cohort taking the training improved very little, with greater improvements from pretest to post-test observed for the second and third cohorts, and the largest improvement for the fourth and final cohort. To produce this grap, first, in the RESULTS window, type SGEDIT ON. Next, run the following code

```
Ods listing sge = on ;
Ods graphics on  ;
proc glm data = plots ;
       class TestType cohort  ;
       model z_total = TestType cohort TestType*cohort ;
       where group = "EXPERIMENTAL" ;
```

The default graph produced has lines that are very light and difficult to see in a presentation. The above code produced graphs in a format editable by the SAS Graphics Editor, a .sge file type. I double-click on this file in the RESULTS window and an editable graph pops up. I right-click in the plot area and select PLOT PROPERTIES. From PLOTS, I select LINE and drag down to the thickness that I want. I want to upload this to my blog, so, from the Graphics Editor FILE menu, I select SAVE AS and save it as a .png file.

And yes, the TestType*Cohort*Group interaction (F=5.84, p < .0001) and the TestType*Group interaction (F=22.92, p < 0001) in the repeated measures ANOVA I did prior to this were significant. Thanks for asking.
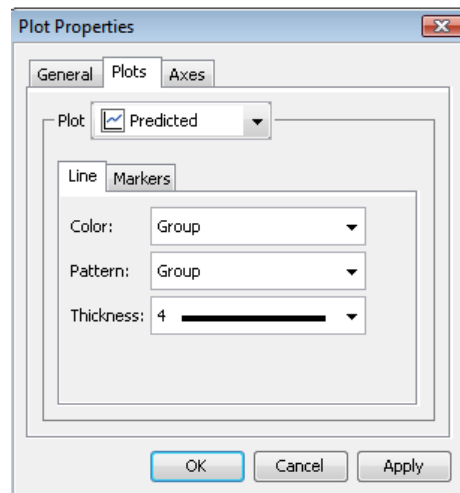
EXAMPLE 2: ELECTION DAY

 The second example uses 2008 ABC/ Washington Post election poll data available from the Interuniversity Consortium for Political and Social Research to map Obama and McCain supporters by geography, region, race and income. What is a bad visualization for one question can be good for another. An example we use is a map of likely voters. To forecast an election, this is misleading, because large states like Wyoming have fewer electoral votes than tiny New Jersey. However, if your question is which regions supported the Democrats, a map is a good choice.

Creating common maps, such as U.S. states or counties, is made easy when you merge the response dataset with the map dataset that is included with SAS 9.2 . The variable I want to plot is the percentage of respondents who selected one out of two choices. There is a STATISTIC = option that you can use for the CHORO statement, but it does not have the option of using the largest category in a frequency distribution.

One simple way to do this is to use PROC TABULATE and create an output dataset. An advantage of using PROC TABULATE is that it also provides me a table of supporting data if I might want to include it

in an appendix or examine the results in more detail. The output file will have two rows for each state. The first is the percent who stated they would vote for Obama if the election was held today (in July 2008). The second row is the percentage of respondents who said they would vote for McCain. Both the summary dataset and the maps.us2 dataset are sorted by state and then merged.

```
PROC TABULATE DATA= in.VOTE2008 OUT=SummaryVOTE2008  ;
        CLASS question3 state ;
        TABLE state, question3*  RowPctN         ;
proc format ;
        value vote
              50.01 - 100 = "Obama"
              0 - 50 = "McCain"  ;
TITLE1 "Likely Vote in 2008 Election";
TITLE2 "ABC/ Washington Post Poll";
PROC GMAP  DATA = SummaryVOTE2008 map = maps.us ;
        ID state ;
        CHORO PctN_01 / discrete LEGEND=LEGEND1 ;
        Pattern1 c = red ;
        Pattern2 c = blue ;
        Label PctN_01  = "Likely Vote" ;
        format PctN_01  vote. ;
        Where state not in (2,15,72) ;
Footnote "Data Not Available for Alaska or Hawaii" ;
```

Points to remember:

The PROC FORMAT step, combined with the FORMAT statement, will show "Obama" or "McCain" instead of a percentage in the legend. More importantly, when combined with the DISCRETE option in the CHORO statement, this formatting will result in all of the states being forced into one of two choices.

The ID statement uses the _map_geometry_ variable that was merged in from the maps.us2 dataset to identify the location on the map.
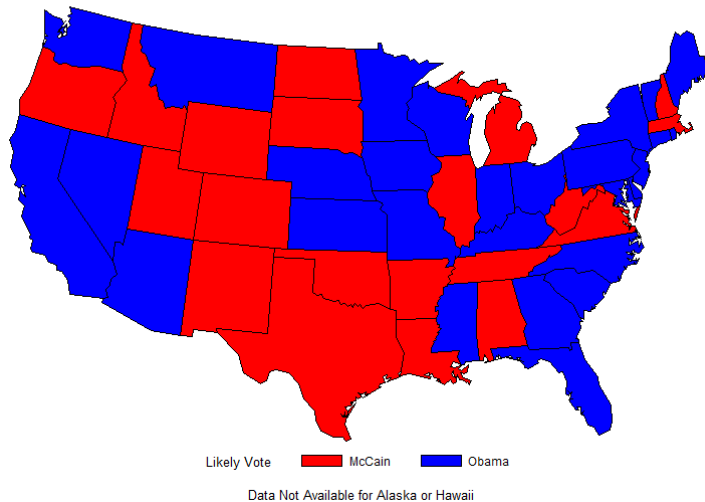

CHORO PctN_01 will plot the variable PctN_01. This is a little trick. Notice I mentioned above that the dataset produced by PROC TABULATE includes two rows for each state, the percentage who would vote for Obama and the percentage who would vote for McCain? The CHORO statement will use the first observation and ignore the others. In effect, it plots the percentage who chose Obama.

At least in July, 2008, it appears that, in terms of geography, support for Obama was predominantly a bi-coastal phenomenon. What about income?
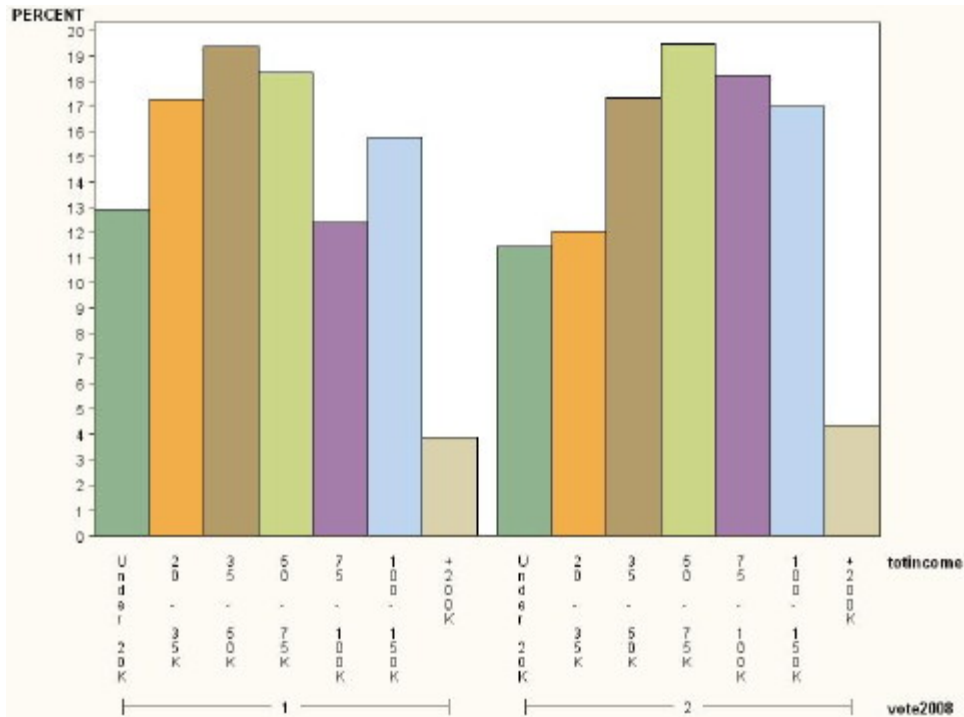
Is there or isn't there? When relationships are marginal is when visual data are least useful. If you have a scatter plot that is just a circle or two histograms that are noticeably skewed in opposite directions, then yes, every picture may be worth a thousand words. You can compare the income distributions of McCain voters side by side using every type of bar chart and pie chart (believe me, I tried) and still no clear picture emerges. Because the picture isn't clear. Using the Mantel-Haenszel chi-square of 3.63, df = 1, probably the appropriate test since it assumes an ordinal relationship, the p value = .0565. The Pearson chi-square, which does not assume ordinality, has a p-value > .20. So, if there is any association at all, it is at best marginally significant and small.

The only code used in producing the above bar chart was running a PROC FORMAT to create the values for the X axis. The graph was actually done wholly with SAS Enterprise Guide by selecting from the TASKS menu GRAPHS and then BAR CHART. A GROUPED COLORED bar chart was selected. A filter for only choices of candidates 1 and 2 was selected under DATA. Under ADVANCED, PERCENTAGE WITHIN GROUP was selected as the statistic to plot.

If income doesn't seem to be a factor, how about race and ethnicity? Let's move from single variable plots and try to look at two variables at once. Rather than look at the respondent's race, I decided to look at the race of their community. I created a variable that was the percentage of African-American and Hispanic voters in the respondent's district. Running PROC UNIVARIATE, I found a mean of 17, mode of 1 and median of 10 for this variable.
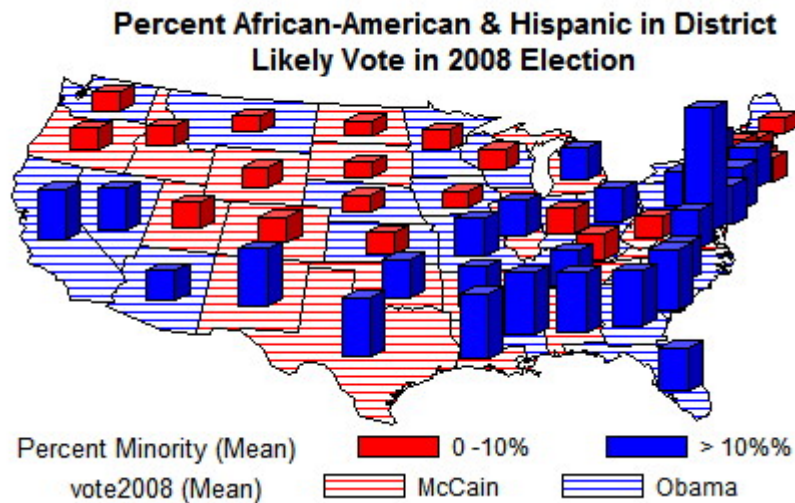
The PROC FORMAT, shown below, defines the values to be graphed. As the distribution was extremely positively skewed, I chose to split at the median, rather than the mean. Note that in this example if the vote was split at exactly 50% it was shown as voting for McCain. Simply by varying these formats, I created several versions of this map with different assumptions and categories, showing the 50-50 states for Obama, or as undecided. I split the percentage minority into three groups. This is an important point we'll return to shortly.

```
proc format ;
      value voten
            0 - .50 = "McCain"
            .501 - 1 = "Obama" ;
      value rangep
            0 - 10 = "0 -10%"
            10.01 - 100 = "> 10%%" ;
```

The first statement below uses the maps.us dataset to map data from my response dataset (a fancy way of saying the file that has my data in it). The AREA statement will color the area of each state based on the value of the vote2008 variable, the option STATISTIC = MEAN is invoked, so the mean for each state will be graphed. Again, the DISCRETE option will force the states to be colored red or blue and not give me 40 different colors based on the actual mean.

The BLOCK statement charts the pctmin variable. The height of the block will be based on the value of the variable, but the color will be determined using the format specified.

```
PROC GMAP   DATA = wuss map=maps.us ;
     ID state ;
     area vote2008 / discrete statistic = mean ;
     block pctmin / discrete statistic = mean  ;
     Pattern1 c = red ;
     Pattern2 c = blue ;
     format pctmin rangep. vote2008 voten. ;
```


Percent African-American & Hispanic in District
Likely Vote in 2008 Election

Percent Minority (Mean) ███ 0 -10%   ███ > 10%%
vote2008 (Mean) ▭ McCain   ▭ Obama

Is there a relationship? It seems pretty clear that there is. In fact, the mean minority percentage in districts where Obama voters live is 21% versus 13% for McCain voters (t= 5.73, p < .0001). Let's talk about what's in the parentheses for a moment. Recently, there has been considerable discussion on the LinkedIn Research Methods and Analytics group surrounding the topic of objections to quantitative methods. The general consensus seems to be that clients/ consumers are often skeptical of our statistical conclusions, sometimes with good reason, sometimes not. Saying that t = 5.73, df = 874, p < .0001 means something to us, but not to them.

On the other hand, if by changing a parameter I can re-run the analysis and show "See, when I split at the average instead of the median I still find this relationship. Now, when I break likely voters into three groups, including undecided, instead of two, I still find this relationship ... "

If I do that five or six times in five or six different ways, I am going to have convinced more people of the robustness of my results than by explaining confidence intervals. To change policy, it's not enough just to be right. You need to convince people you are right and sometimes that means showing them – literally.

EXAMPLE #3 VISUAL DATA WITH JMP
There are two reasons to use JMP.

One is if, like me, you prefer to use a Mac but much of your day is spent analyzing SAS datasets. If you have set up a SAS metadata server, you can open a SAS dataset from JMP. If you haven't set up a metadata server, don't have a SAS license, or if you are some place, say, Tunisia, where you cannot access your SAS server, if you have saved the SAS dataset as a .xpt format, you can just open it from the FILE menu.

The second reason to use JMP is it is an extremely simple way to get an astounding array of high quality graphics very easily.

For this project, I needed to predict which high school students were interested in joining the military. This is not a trivial question. The U.S. military is the largest organization of ANY KIND in the entire world. It is by far the largest percentage of the federal budget. Last year, the army alone spent $216 million on advertising and 95% of recruits have a high school education, so this is the population being targeted.

Data came from the Monitoring the Future study of 2,596 students enrolled in the tenth-grade in American high schools in 2008. The dependent variable we are interested in modeling is whether the student's post-secondary plans include joining the military (probably or definitely will) or not.

Decision Tree, ROC Curve and Lift Curve from a SAS Dataset by Clicking

Yes, I know it sounds like sacrilege.

A decision tree is produced in JMP by selecting ANALYZE > MODELING > PARTITION and then identifying the Y variable, in this case, a dichotomous variable indicating interest post-secondary plans for military service, and the X variables of sex, mother's education, father's education, maternal employment, hours worked, grades, expect to attend a four-year college, if the student had ever been suspended, failed a grade or attended summer school, how much time student spent home alone each week, attendance at summer school, cigarette smoking and how often student talked to parents. In short, variables measuring socioeconomic status, academic success, social problems and family environment were included. To begin the decision tree, simply click on the SPLIT button. A decision was made to only include variables that increased the R-squared by at least 1%. With these four branches, the R-square was 9.3%,
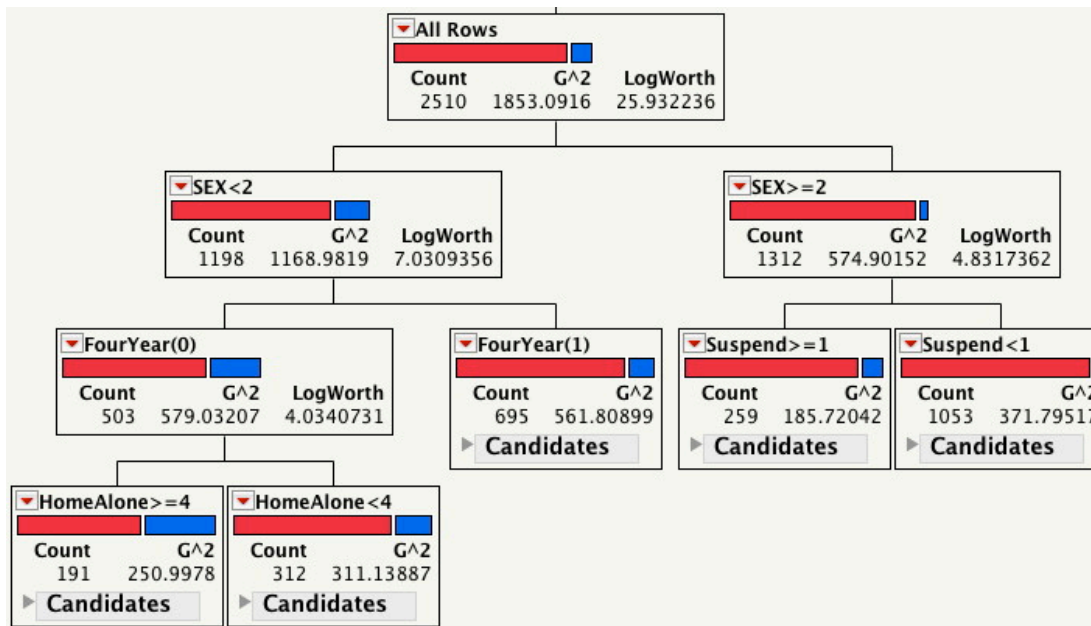
Speaking of re-running analyses .there are several advantages to JMP in presenting results and addressing the "You can prove anything with statistics" argument. First, it is very simple to select out a random sample of rows to use to "train" and create an equation using those rows. Just select the rows , right-click in the first column (what would be the OBS number for those of you more familiar with SAS) and select "EXCLUDE/UNEXCLUDE". The selected rows will be excluded from the analysis. The top of your decision tree will show two values for R-squared, for the records included in the computation of the prediction equation, and for the excluded variables. You can do this for many JMP procedures, it isn't limited to partitions. Of course, the R-square for the excluded group is generally lower, but if it is not much lower, it does lend some credence to your argument that your equation is useful.

This graph simply colors the points and shows the proportion for each split. You can see that over a third of the males (37%) who are home alone often and who don't think they will attend a four-year school say that they plan to attend the military after high school. Since the percentage for the whole population was about

12%, this is a substantial increase. If you were creating a military recruitment campaign, you might want to think about having ads on TV programs or websites with a predominantly male audience, running between 3 and 6 pm when students are most likely to be home alone.
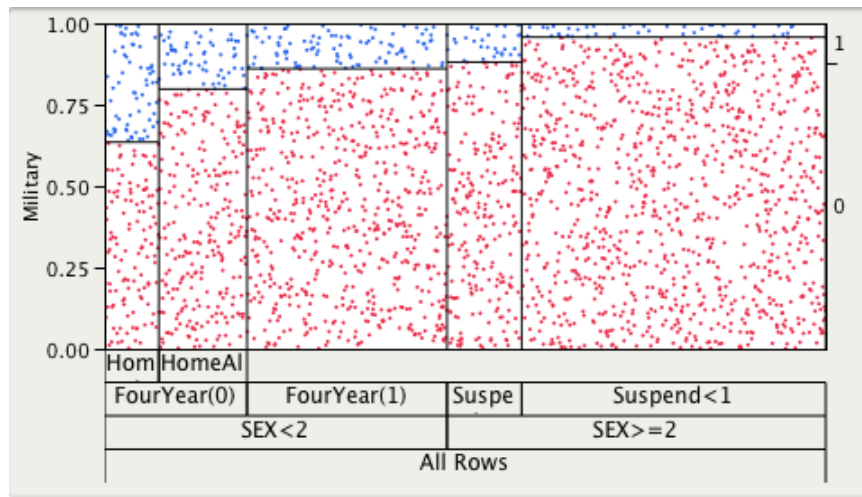

How stable are these results? As statisticians, you're very familiar with the idea of multi-collinearity. If you have correlated independent variables, as many of these are likely to be, you may get a very different outcome if you run the analysis again. Obviously, you test for multi-collinearity by examining the Tolerance or Variance Inflation Factor. Obvious, to you. Having tested for multi-collinearity, as well as examined the correlation matrix of independent variables, I proceeded to run this analysis several more times, using slightly different assumptions or slightly different variables, e.g., the average of mother and father education rather than each entered individually. Each time, I also selected a new training sample. This is similar to boot-strapping, of course, where one re-samples from the same population, but not exactly, as I'm also sampling from the domain of possible independent variables simultaneously.

While discussions of boot-strapping to obtain estimates and confidence intervals for standard errors and VIFs may not be convincing (or even penetrable) to the non-technical audience, running the same analysis with minor variations and producing essentially the same results several times in a row, does lead to the conclusion that (not surprisingly) gender makes a difference her. However, so does whether you've been suspended from school and how much time you spend alone. This brings up the question of why students are alone. Are these students more likely to be from single-parent families? We don't know because that was a variable we did not have in the dataset. Might it be important? Maybe, that would be a good question to follow up in a subsequent analysis. What is happening here is that we're moving from a monologue on the results to a discussion.

The next point to discuss is trade-offs in decision-making, and this is where the Receiver Operating Characteristic (ROC) curve comes in useful. [Click on the red arrow at the top left of the partition window for pull-down options include ROC and Lift curves.] The most familiar tool for most statisticians who are predicting a categorical dependent variable is logistic regression. In logistic regression, we find the set of b coefficients that have the maximum likelihood of producing the observed data. That sounds like a good thing and usually it is. There is a problem, though. The LOGISTIC procedure fits a model based on the cumulative slopes of the response probabilities. However, sometimes we think that one response category is more important than another. While neither JMP nor the ROC curve solves this problem, what the ROC curve does do is allow us to make choices, to decide where we want to make these trade-offs among the probabilities.

In an ROC curve, the true positive rate (Sensitivity) is plotted against the false positive rate (100 – Specificity). One of the beauties of this curve it is simple to explain. You can see that you're always going to have zero false positives if you have zero true positives. Let's say we predict no one will join the military. Well, we will have not falsely predicted for anyone, so that's good, but we also did not identify any of the people who would join (no true positives). On the other end, if we predict everyone will join, we'll correctly identify everyone who intends military service but we'll incorrectly identify 100% of those who don't enlist.

Miltary Recruitment Prediction	A "Better" Model
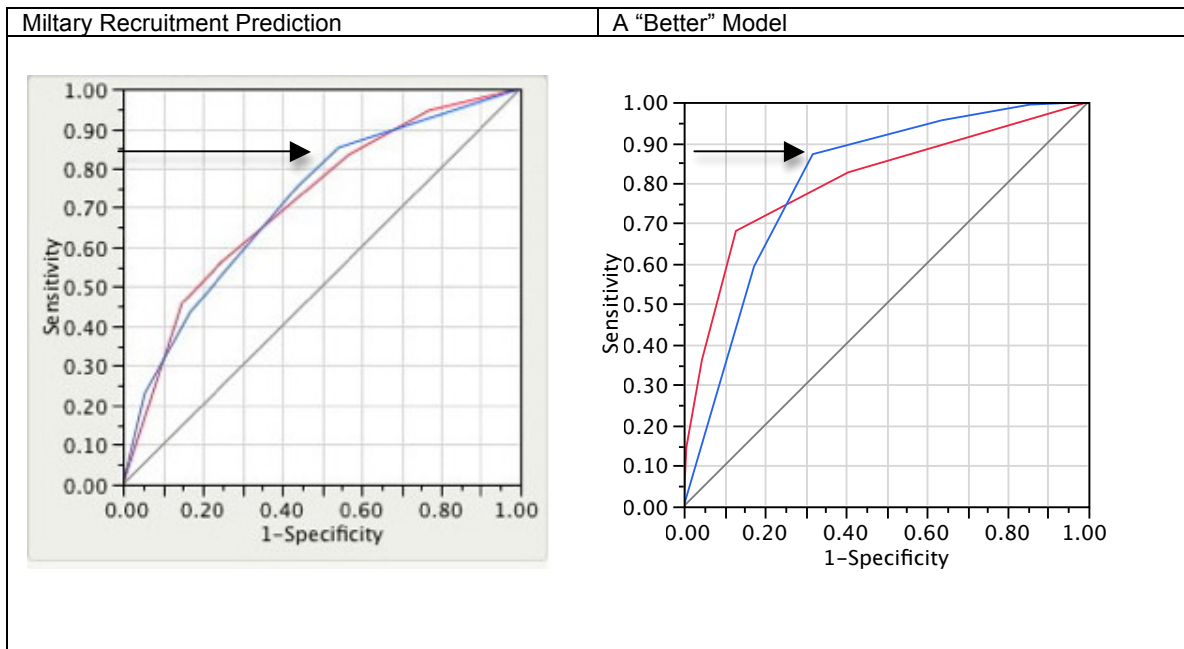
Two models are shown. The one on the left was the best model we could find with the set of predictors available. Let's assume that we have made the decision that we need to identify correctly at least 85% of the recruits, that is 255. To get that level of "true positives" we'd have to accept 55% false positives, or spend money trying to recruit 1,210 people (.55 * 2,200). Since each recruit is so valuable, we'd be willing to market to 1,465 people to meet that target.

On the other hand, our better model allows us to get that same sensitivity rate with only 30% false positives. With our better model, we can get the same number of recruits by marketing to only 915 people. Unfortunately, this model is a hypothetical one I just made up. If this had been a real model, this would be the point for discussing whether the cost of getting the additional data outweighed the cost of marketing to 550 more people. Since it isn't a real model, the question is, what variables did we not include that might get us to this level of prediction? Family history of military service? Race? Income? Geographic region?

Conclusion

Someone once said that a statistician is a person who was good at math but didn't have enough personality to be an accountant. Personally, I think we are awesome, but I have experienced enough business meetings (and three teenage daughters) to realize that opinion is not universal.

Bessler, in a SUGI 25 paper asserted that while people make quicker decisions with graphs they make more reliable decisions with tables, and thus the best choice of presentation may be a graph and table combined. That depends on your purpose. In my case, as an evaluator, for example, I have already made the decision backed up by considerable tables of statistical analysis. My challenge now is to tell the story of what I found in a manner that informs the non-statistician executives or other experts sufficiently that they accept the conclusions and support the decisions based on the data.

What I am NOT saying is that we need to discard measures such as standard errors, R-square, prior probability and F-values. However, as we move into an era of more team science it is crucial that we leverage every tool we have to convey meaning to our audiences. It is not enough that the data and conclusions make sense to us. We need to use whatever tools we have available – and visual data is a great one – to make sure that our findings reach, involve and convince people beyond just us. In doing so, we can turn a presentation into a conversation. And that just may be an art after all.

REFERENCES

Bessler, L. (2000) Show Them What's Important:Solutions for a Finite Workday in an Era of Information Overload. Paper presented at the annual meeting of SAS Users Group International.

Steele, Julie & **Ilinsky**, Noah. (Eds.) Beautiful visualization. Looking at data through the eyes of experts.. Sebastopol, CA: 2010.

CONTACT INFORMATION

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

Name: AnnMaria De Mars
Enterprise: The Julia Group
Address: 2111 7th St #8
City, State ZIP: Santa Monica, CA 90405
Work Phone: (310) 717-9089
E-mail: annmaria@thejuliagroup.com
Web: http://www.thejuliagroup.com