Using the JMP® Statistical Package to Enhance the Statistical Literacy of the U.S. Energy Information Administration's (EIA's) Survey Staff

Carol Joyce Blumberg[1], Marlana Anderson[1], Andrew Hoegh[2] and Jennifer Lawhorn[3]

[1]U.S. Energy Information Administration, 1000 Independence Ave. SW,
Washington DC 20585. Email: carol.blumberg@eia.gov
[2]Virginia Tech, Department of Statistics, Hutcheson Hall, Blacksburg VA 24061-0439
[3]Edwards Lifesciences, M/S JAM5, One Edwards Way, Irvine CA 92614

Abstract

EIA began using JMP® five years ago so that survey managers could do some statistical analyses on their own without having to write SAS® code (or depend on others to write it.) EIA started with 10 copies. As word spread of the package's ease of use, the number of users increased incrementally to 50 within three years. The users have a wide variety of backgrounds in terms of knowledge of mathematics and statistics and in the use of statistical packages. A series of workshops was developed using EIA mathematical statisticians as the instructors. The initial intent was to concentrate on how to use JMP® and include some short discussions on how to read the output. However, each time the workshops were given the attendees had more questions about the mathematical and statistical concepts than on how to generate output. These questions ranged from how to read exponential notation, to how to read a stem and leaf plot, and to why do statisticians prefer "do not reject H_O" to "Accept H_O". This paper will focus on the types of statistical and mathematical questions asked by the workshop attendees and the variety of methods used to answer these questions.

Key Words: Statistical literacy   Software use   Survey management   Training

1. Introduction

By 2007 it became apparent to the management in the former Office of Oil and Gas at EIA that there was a need to have a statistical package in addition to SAS®. There were several reasons for this: 1) the number of statisticians available to help other staff had been reduced; 2) the survey managers wanted to be able to do some statistical analyses themselves; 3) statisticians, survey managers, and other staff wanted an easy-to-use computer package that was able to produce presentation-quality graphical displays; 4) several staff had seen JMP® used in presentations they had attended outside of EIA for creating graphical displays and were very impressed; and 5) many universities and colleges were replacing SAS® and other similar packages (e.g., Minitab and SPSS) in their statistics courses by JMP® and R. The first author had been a faculty member at Winona State University until 2006 and had used JMP IN® while teaching undergraduate statistics courses (including introductory statistics) there. Hence, she was asked by her supervisor to investigate whether it would be advantageous to some EIA staff to use JMP®.

As an infrequent user of JMP IN®, the first author did not have very high expectations for JMP®. She was given a 30-day free trial access to JMP®, as can anybody interested in exploring JMP® (see http://www.jmp.com/ads/adwords.shtml?gclid=CLuM_5uth68CFYGo4AodLhn6-Q for details.) Immediately, she was impressed by the ease of use and how similar it was in many ways to Excel. Since Excel was used regularly by most EIA staff, this was important. It was also quickly apparent that JMP® could easily read and produce Excel files. During this 30-day period, two staff from the JMP® office in Delaware came to EIA and discussed the implementation of JMP® in a Federal Government setting with the statisticians located in the Office of Oil and Gas. It was also pointed out that, at the time, the per copy cost of JMP® under a Federal Government contract was only about one tenth of the cost for SAS®. Hence, the decision was made to purchase 10 copies of JMP®, with 4 copies going to statisticians (including the fourth author—who was an intern at EIA during at the time) and 6 copies going to survey managers.

However, once JMP® was loaded on to the 10 machines, some training had to be done of both the statisticians and the survey managers as to how to use JMP®. In addition to learning the JMP® package, the survey managers also needed to learn how to interpret the most common types of output that they would probably generate using JMP®. The purpose of this paper is to explain the initial training that was done and later training as EIA increased the number of copies of JMP® from 10 to 30 and then to 50, as users told others about their use of JMP®.

2. Workshop Materials and Format of the Workshops

The first time the workshop was taught, a preliminary outline was developed by the first and fourth authors. It was sent to the 8 people who would attend the training on JMP®. Based on their comments, the outline was modified. This same procedure of sending out an outline and getting back feedback was also done for the second and third workshops. An example outline is shown in Appendix A.

The first time the workshop was presented, a number of handouts were developed. These were of two types—technical step-by-step instructions and sample outputs. An example of a sample output is attached as Appendix B. At the final meeting the first time the workshop was given, the suggestion was made to put all of the technical step-by-step instructions into a users' manual. The people attending the workshops also suggested additional topics for which they wanted step-by-step instructions and these were also included in the users' manual. A copy of the Table of Contents of this manual, called "Introduction to JMP® for EIA Users", is attached as Appendix C. This "Introduction…" plus sample outputs served as the course materials the second and third times the workshop was given. As with the first time, additional sections were added to the manual based on suggestions from the workshop participants. Also, as new versions of JMP® were obtained by EIA, the manual was updated to reflect changes in the new versions. As can be seen in Appendix C, while most of the topics covered are applicable to all JMP® users, some topics are specific to EIA. For topics that are not EIA-specific, examples are chosen, whenever possible, that are EIA-specific or mirror the features of data sets that are analyzed or produced by EIA.

Each time the workshop consisted of three sessions of 90 minutes that were held approximately two weeks apart. This was done for two reasons. First, the survey managers have weekly and monthly production schedules that are very tight. The sessions had to be fit around these production constraints. Second, by having a few weeks between the sessions, this allowed the new users to be able to practice on their own and suggest topics for the next sessions when they could not figure out how to do something. The instructors were also available between sessions to help people individually.

The workshop sessions were very informal. In addition to the specially-prepared handouts, a computer and projection system was available to allow access to JMP® during the sessions. The leading of the sessions was split each time between the two instructors, with the first author covering the basics of JMP® and some of the mathematical and statistical procedures. Other statistical procedures, graphics and scripting were presented by the other authors. One important feature of the sessions was that people were very willing to ask questions. Thus, what was planned was often not what was covered (especially the first time). The next section of this paper will explain some of the basic statistical literacy questions that were asked and how answers to these questions were developed.

3. Some Examples of Statistical and Mathematical Literacy Questions and Answers

One of the hardest things to do in planning and doing the workshops was to come up with materials that addressed the needs of the variety of JMP® users, who ranged from survey managers with minimal mathematics and statistics background (of having only one mathematics and/or statistics course in college) to Ph.D. statisticians. The examples presented in this section of the paper were obviously needed more by those with weaker backgrounds. It is important to point out, however, that those with the weaker backgrounds are very bright, highly-valued and well-respected EIA employees.

Exponential Notation on Data Tables

One of the sample outputs handed out involved some data with very large values in it. For example, in a JMP® data table where the width of a certain column had been set to 4, the actual value of 32396 on the printed copy of the table taken from the Internet and handed out with the sample output, appeared as 3e+4. One of the survey managers commented that she had seen that notation on her calculator and a few other places and she then asked for an explanation of what it meant. The answer of “It is scientific notation” was not sufficient, since the person remembered seeing scientific notation many years ago, but did not remember much more about it. The first author then explained scientific notation for several examples that included both large numbers and very small positive numbers. Also, since JMP® appeared to give $3.0000\text{x}10^4$ instead of $3.2396\text{x}10^4$ (even though the complete value appears if the user double clicks on the cell in the data table while in JMP®) the survey manager then asked about this. Others then chipped in and there was an interesting discussion about rounding in general and when it is appropriate and not appropriate.

Rational Exponents

One of the features of JMP® that can be confusing to users is its use of rational exponents. It has both an $x^y$and a $\sqrt[y]{x}$ function. So, if people want to compute the cube root of a variable (say, v1), as in high school mathematics, they have two choices, they can either use, $\sqrt[3]{v1}$ or $(\text{v1})^{1/3}$. As expected, all three times there were people who remembered seeing rational number exponents, but did not remembering how to use them. Some examples of how to compute quantities involving exponents in both JMP® and Excel were shown to the workshop attendees. Further, the presence of two ways to handle exponents (and that the two ways used reciprocals (that is 3 versus 1/3) as the information that the user had to enter was very confusing to those with less mathematics background. Since this came up in close proximity to the discussion of the "e" notation the first time the workshop was given, the psychological phenomenon of interference occurred for some of the people in the workshop. Therefore, an explanation was unexpectedly needed of the difference between a (Number1)x10$^{\text{Number2}}$ and (Number1)$^{\text{Number 2}}$.

Stem and Leaf Plots, Extreme Values, and Rounding/Truncating

A nice feature of JMP® is that it is easy to get stem and leaf plots. Since Excel does not do stem and leaf plots and in SAS the default plot in PROC UNIVARIATE in the past was a histogram (horizontal bar chart), each time the workshop was given, stem and leaf plots were presented to the workshop participants. In each of the workshops there was at least one person who had seen a stem and leaf plot, but freely admitted to not knowing how to interpret one.

Hence, an appendix about box plots and stem and leaf plots was included in the "Introduction to JMP® for EIA Users" guide. This appendix was part of a chapter written by the first author (Blumberg, 1990) for inclusion in a booklet for secondary teachers and students developed by the ASA Section on Statistical Graphics. The booklet was purposefully not copyrighted. A copy of the booklet chapter is available from the first author.

Because the appendix explained how to construct a stem and leaf plot by hand, the emphasis in the workshops was on how to interpret a stem and leaf plot. The first example given in the workshop was a very simple one with just a few numbers that involved no rounding and had a legend that was easy to interpret. The second example used was the stem and leaf plot on the second page of Appendix B of this paper, which is summarizing the BTUs (in thousand megawatt hours) generated in 2010 via hydroelectric power for the 50 States and the District of Columbia. This stem and leaf plot was used because of its ambiguous legend of "0|0 represents 0". This legend is even confusing to those who have seen stem and leaf plots in the past.

Hence, the important point that one should go back to the raw data to determine the meaning of a graphical display was able to be made here, since the only way to determine the digits in this stem and leaf plot was to look at the printed data table. By looking at the printed data table, the workshop participants could see several things. The instructors had eliminated some of the extremely high values for the purposes of getting a nice example for the purposes of the workshop. In one of the times that the workshop was presented, this led to a short discussion, with the statisticians disagreeing among themselves, as to when and when not to delete extreme values. Each time the workshop was given, by looking at the printed table, the non-statistician workshop attendees quickly realized that it was not easy to tell exactly what was going on in the stem and leaf plot because the first few States alphabetically had low BTUs. With a bit of prodding, they figured out that the top value of 9|2 represented 923 and so for all values in the stem and leaf display, the stem represented the hundreds place and the leaf represented the tens place.

However, the value of 9|0 was more problematic to several people each time. There were no values in the printed table in the 900 to 999 range. But, the 9|0 must still represent the second highest non-extreme value, which turned out to be 898. Hence, the workshop participants could conclude that JMP® must be rounding the numbers. But, some people in the workshops had been taught to truncate when making a stem and leaf plot and others had been taught to round. This led to a short discussion of when it was appropriate to truncate and when it was appropriate to round when presenting data to the public.

Changing the Axes on Scatterplots in both JMP® and Excel

The second and third times that the workshops were held, the JMP® "tool" of Graph Builder was discussed. One very nice feature of this tool is the ability to easily change the axes of scatterplots (and histograms and other graphics) by the use of sliders or a drop-down menu. Since much of the data analyzed at EIA is skewed and/or contains some extreme values that are outliers (e.g., the BTU example of the previous subsection), this ability to easily focus in on certain portions of the data is very useful to both survey managers and those doing analyses (from either a statistics or economics point-of-view). It was only the third time that the workshops were held that the topic of adjusting the axes was discussed in detail. This also led to a discussion of how to make scatterplots and line graphs in Excel more readable in terms to adjusting the axes. The idea of focusing in easily on certain portions of the data has been used subsequently by several of the workshop attendees to better explore their data and relationships between variables.

Data Visualization

Almost all of the workshop attendees were not used to dynamic data visualization. In addition, the variety of graphical and data visualization techniques known by the participants was highly variable. One of the nice features of Graph Builder is that it is highly dynamic. This use of dynamic data visualization, other than already described for use with the scatterplots, was fascinating to many of the attendees. In fact, several attendees have used these dynamic features since the workshops and even used graphical techniques they had not seen previously.

Logic of Hypothesis Testing

While the main focus of EIA weekly and monthly published data is on point estimates, some attention is given to standard errors. However, little attention is paid to hypothesis testing. The statisticians and econometricians, of course, automatically make the connection between standard errors and hypothesis testing. But, the survey managers do not do so as easily. The survey managers sometimes overemphasize non-significant differences (both in terms of statistical non-significance and meaningful non-significance). Most are older and the statistical training they had in their college studies was hand- and calculator-based. Even though Excel can do some basic hypothesis tests, most of the survey managers are unfamiliar with that capability in Excel. So, since hypothesis testing in JMP® is much easier than in Excel, it was decided to include interpretation of hypothesis testing output as part of the workshops each time. When reminding those in the workshop of the logic of hypothesis testing, the first time the workshops were offered one of the participants asked the question of what exactly the phrase "Do Not Reject the Null Hypothesis" meant. In answering the question, the distinction between the wrong phrasing of "Accept the Null" was discussed. One effect of this was that the survey managers who took the JMP® workshops still sometimes over-interpret non-statistically significant differences, but now are much more careful, on the whole, about over-interpretation. The second and third times the workshops were offered the participants were given a handout, from a textbook that is no longer published, that nicely explains the logic of hypothesis testing. Several of the workshop attendees mentioned that this handout was very useful. It should be noted that the first author received permission from the publisher to make a handout of this section from the textbook as well as the section that explains probability values.

Probability Values

One nice feature of JMP® is that, when reporting the results of a hypothesis test, it gives a picture (see second page of Appendix B for an example) interpreting the p-value. Those attending the workshops found this picture very confusing at first. However, having it on the output was a huge advantage to the instructors when trying to explain probability values, because the pictures were in the same format as those in the handout on probability values, even though the handout was about 20 years old. This helped remove some of the mystery about probability values.

The output also gives three p-values: the two-tailed one and the two one-tailed ones. This is similar to how Excel gives p-values. Having three p-values given was confusing to the non-statisticians at the workshops. But, having all three p-values on the output along with the picture actually helped those who had previously not understood p-values to gain a basic understanding of p-values.

4. Benefits of Using Co-Instructors of the Workshops

The first author had taught at the college level for 33 years before coming to EIA in 2006. The second to fourth authors all had recently been in graduate school (or still attending graduate school) when they helped with the workshops. Hence, the first author’s skills with interactive software, such as the Graph Builder portion and Scripting feature of JMP®, were much more limited than those of the second to fourth authors. Without having the second to fourth authors as co-instructors, these important interactive features of JMP® would have not been properly explained to the workshop participants. Also, all three co-authors were wonderful sounding boards for ideas about organization and presentation ideas, including making excellent suggestions of how to explain certain topics and examples to use.

The benefits for the second to fourth authors were quite varied. One obvious benefit was resume building. Another huge benefit was that they were able to interact with EIA employees outside of their workgroup, which gave them greater visibility within the organization. In addition, it gave them an opportunity to reinforce knowledge they had learned in their statistics courses.

In terms of the teaching aspects, it allowed the second to fourth authors to observe EIA employees with different backgrounds and amounts of knowledge about statistics. It was a great lesson for them in the cross-functionality that always accompanies a statistical team, other groups and the interactions between statistical teams and other groups. It also helped them to see how those who had limited statistics training might view certain statistical functions and terms, such as a stem and leaf plot or standard deviation. In particular, they realized that sometimes people have seen these terms, but may be misinterpreting the meaning of the terms and the results that depended on them. Finally, two of these authors had minimal teaching experience and found co-teaching an interesting experience.

A benefit of the workshops for all four authors was learning about some of the types of analyses that the different groups within EIA used on a regular basis, especially those analyses that were different from those commonly used by the groups in which they normally worked. It was also very interesting to learn how others at EIA wanted to use JMP® or were already using JMP®. This even allowed the authors of this paper to learn capabilities of JMP of which they were not aware of previously and to then use these capabilities in their own work at EIA. Further, all of the authors agree that a benefit of having to present to others is that they learned the JMP® package better than if they had been just attendees of the workshops.

Reference

Blumberg, C. J. (1990). Simple graphical displays for looking at bunches of data. In B. Gunter (ed.), Seeing and Believing: A beginner's guide to statistical graphics (pp. 25-43). Published by the Statistical Graphics section of the American Statistical Association.

Disclaimer and Acknowledgements

The views expressed in this article are those of the author, and no official endorsement by the U.S. Energy Information Administration or the Department of Energy is intended or should be inferred. The authors wish to thank Janet Gordon and Carol French of EIA for helpful editorial comments.

APPENDIX A

JMP Workshop Tentative Outline

Thursdays, May 5, May 19 and June 2
9:30 to 11:00 a.m. in 2G-024

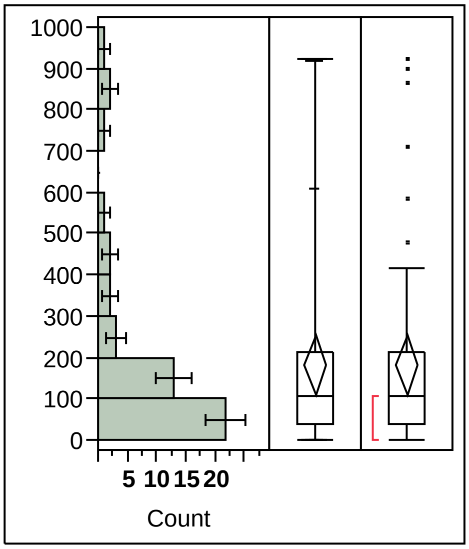
Leaders: Carol Joyce Blumberg and Andrew Hoegh

1. Introductions
2. Approach we will take—informal with handouts and demonstrations
3. Getting started and getting Help
4. Hydroelectric data set from CNEAF and our JMP file of it
5. Graphical displays
    a. Histograms
    b. Stem-and-Leaf display
    c. Boxplots
6. Basic summary statistics for a single variable
    a. Mean, median, mode, and quantiles
    b. Standard deviation versus standard error and confidence intervals
7. Other summary statistics (including skewness, kurtosis, and CV)
8. Saving, formatting, and exporting files
9. Creating data tables
    a. From scratch
    b. From an external file
10. Being careful about the level of data and definition of JMP terms for this
11. Changing the characteristics of a column
12. Formatting of variables that are dates and times
13. Excel-type manipulations
    a. Adding columns and rows
    b. Changing data at the cell level
    c. Formulas
14. Scripts of commands
15. Making subsets of a JMP file
16. Comparing two “variables” where the X variable Is nominal and the Y variable is continuous
    a. Histograms and Boxplots using the Analyze—Distribution menus
    b. Descriptive statistics
17. Graph Builder
18. Comparing two “variables” where both X and Y are continuous
    a. Scatterplots
    b. Regression lines
    c. Confidence intervals and prediction intervals
19. Joining/Merging Files (as time permits)
20. Hypothesis testing involving t-tests and ANOVA (as time permits)
21. Other graphical displays (as time permits)
22. Time-series plots (as time permits)
23. Random number generation (as time permits)
24. Quality control techniques (as time permits)

APPENDIX B
Example of Handout

Numerical Output (without Outliers Greater Than 2000) & With Lots of Trimmings

Distributions
Total 2010



Quantiles

100.0%	maximum	923
99.5%		923
97.5%		918
90.0%		610.6
75.0%	quartile	213
50.0%	median	104
25.0%	quartile	37
10.0%		1.8
2.5%		0
0.5%		0
0.0%	minimum	0

Moments

Mean	184.80851
Std Dev	242.79591
Std Err Mean	35.415423
Upper 95% Mean	256.09606
Lower 95% Mean	113.52096
N	47
Sum Wgt	47
Sum	8686
Variance	58949.854
Skewness	1.9757501

Kurtosis	3.1838465
CV	131.37702
N Missing	0

Stem and Leaf

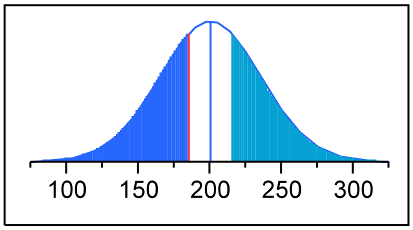
Stem	Leaf	Count
9	02	2
8	7	1
7	1	1
6		
5	9	1
4	18	2
3	34	2
2	145	3
1	00012222345578	14
0	000000112334444566778	21

0|0 represents 0

Test Mean

Hypothesized Value	200
Actual Estimate	184.809
DF	46
Std Dev	242.796

	t Test
Test Statistic	-0.4290
Prob > |t|	0.6700
Prob > t	0.6650
Prob < t	0.3350



Confidence Intervals

Parameter	**Estimate**	**Lower CI**	**Upper CI**	**1-Alpha**
Mean	184.8085	113.521	256.0961	0.950
Std Dev	242.7959	201.7573	304.9482	0.950

Prediction Interval

Parameter	**Future N**	**Lower PI**	**Upper PI**	**1-Alpha**
Individual	1	-309.086	678.7031	0.950
Mean	1	-309.086	678.7031	0.950
Std Dev	1	.	.	0.950

APPENDIX C

Table of Content from "Introduction to JMP$^{®}$ for EIA Users"

Topic **Pages**

The Appendices are located in the JMP part of the EIA Intranet.