

Use and Applications of JMP® in Microbial Ecology

Jose F. Garcia-Mazcorro*, Ph.D.

South Central SAS Users' Group Educational Forum 2012

University of Houston, November 5-6, 2012

*Current address: Facultad de Medicina Veterinaria y Zootecnia, Universidad Autónoma de Nuevo León. Francisco Villa s/n Col. Ex-Hacienda el Canadá. Gral. Escobedo, N.L., C.P. 66050, México. sites.google.com/site/josefgarciamazcorro

Abstract

Microbial ecology is the ecology of microorganisms; that is, the study of the relationship among microorganisms and between microorganisms and their environment. Millions of dollars and countless hours are currently spent by enthusiast scientists, microbiologists and bioinformaticians from all over the world to elucidate the characteristics and biological mechanisms behind the ecology of microorganisms.

Microorganisms are omnipresent in the biosphere. They can be found in virtually every environment, from bright-colored hot springs to the skin and guts of animals. However, perhaps because of their close relationship with our bodies, the microorganisms inhabiting the intestinal tract have received much more scrutiny than any other group of microorganisms on Earth. More importantly, gastrointestinal microorganisms have been widely studied because they strongly modulate health and disease in humans and other animals.

JMP® (SAS Institute) is a powerful software that offers a variety of tools and applications that have been used in many fields of study. Surprisingly, JMP® has only been scarcely used in the field of Microbial Ecology. In this paper, I will discuss the use and applications offered by JMP® 9.0.0 in the study of microbial ecology with a strong emphasis on gut microbial ecology. Topics will include basic statistics, graphical tools, and multivariate methods such as clustering and principal components. Real data sets obtained from animal studies will be utilized to demonstrate the usefulness of the software. Additionally, examples will be given from other fields of study for people outside the field of Microbial Ecology.

Introduction

With few exceptions (Schulz and Jørgensen, 2001), all microbial life on Earth is invisible to the human naked eye. For example, the size of the bacteria living in our gut is about a thousandth of a millimeter. Because of our inability to observe microorganisms directly, for centuries we did not pay attention to the millions of tiny life forms that occupy virtually every environment in our Planet. It was until Antonie van Leewenhoek designed a microscope in 1676 that humans started observing for the first time the vast microbial life. One can only imagine the fascination that these minuscule life forms caused on van Leewenhoek's mind.

What is microbial ecology?

When I talk about Microbial *Ecology* to my students in Mexico, some admit that they did not know that the microorganisms were related to recycling and pollution of the environment. These thoughts represent a general unawareness of what ecology is really about. Ecology is much broader than the mere study of pollution, it is the study of organisms and their biological and non-biological interactions with their surrounding environment. It is about, for example, how the organisms respond and adapt to changing environmental conditions; how, why and when populations increase, decrease, and disappear; how organisms compete for resources; how diseases spread; and why organisms travel long distances to mate and reproduce. Ecology is one of the most interesting fields of study because it is concerned with life and its relationship with their surroundings. Microbial ecology can be thought as a branch of Ecology: it is simply the ecology of microorganisms.

Studying microbes in their natural environment

In contrast to other natural environments, such as a jungle or a desert, one cannot simply go out and observe the behavior of microorganisms. It is just not that simple. Moreover, most microorganisms are virtually the same in terms of size and shape; therefore, even if we would be able to watch microorganisms in their natural environments, we also would not be able to obtain much information about them. At most we would observe a bunch of tiny rods and spheres replicating asexually at regular intervals of time. These are some reasons why scientists have developed many different methods to study who the microorganisms are and what they do.

The first thing we would need to know to study microorganisms is their name, but naming microorganisms has proven to be one of the most challenging things to do. No many years ago, it was common practice for Microbiologists to obtain natural samples, for example soil, water or intestinal contents, and try to grow artificially the microorganisms within. For example, in 2002 one particular microorganism was isolated from human feces and was found to produce butyrate, formate and lactate, but no hydrogen as fermentation products (Duncan et al., 2002). Based on its place of origin, this microorganism was proposed to be called *Faecalibacterium faecalis*: pertaining to feces; *bacterium*: a small rod). Interestingly, today we know that different types of *Faecalibacterium* play a very important role in intestinal health of humans (Sokol et al., 2009) and dogs (Suchodolski et al., unpublished). Anybody could see that this approach of naming microorganisms is completely arbitrary in that it does not follow any natural law. Indeed, this classification system is possible but the problem is that microorganisms are too many to be named like that, they live in every possible environment on Earth, and are difficult to grow artificially because we simply are unable to mimic their natural conditions.

If naming microorganisms has proven to be difficult, the study of their biological activities represents a different level of complexity. Microbes produce and consume numerous substances in their natural environments, although these substances can fortunately be detected with precision using analytical instruments.

Computers and software to study microorganisms

Microorganisms are very tiny and very numerous. A single gram of soil may contain up to 100 million bacteria (Whitman et al., 1998), more than 10 times the number of people currently living in Mexico city! Like in many other fields of study, Microbial Ecologists had to turn to computers and software to analyze and store data from microbial communities.

The data that comes out from microbiological studies can be quantitative or qualitative. For example, researchers have counted the number of microorganisms in a given sample, such as intestinal contents or soil, they have also measured the volume of certain substances produced by the microorganisms, or have grouped microorganisms based on a variety of characteristics, such as biochemical behavior or genetic relatedness.

What is JMP®?

JMP® is a visualization and statistical software created by SAS co-founder John Sall in 1989. Over the past 23 years, JMP® has evolved into a broad family of statistical discovery products used in many industries around the world. Today, John Sall remains the chief architect of JMP® software (www.jmp.com/about/).

Using JMP®

As in any other software, JMP® requires you to get familiarized with the different options and tools that are not necessarily obvious for you to come and pick. The top part of the JMP® home window is presented in **Figure 1**. The tabs on the top are File, Tables, DOE (Design of Experiments), Analyze, Graph, View, Window, and Help, which can easily lead you to your main objective. Unlike most software, JMP® includes a comprehensive literature (including entire books!) in the Help option. In addition, there are a number of icons that will lead you directly to your desired objective (**Figure 1**).

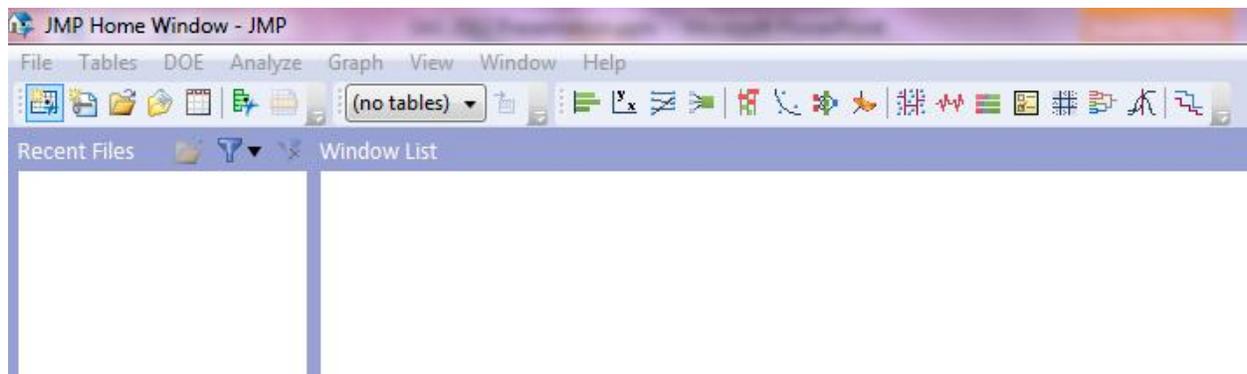


Figure 1 Top part of the JMP Home Window.

Once you open a file, you can get access to it by just clicking the title in the recent files window (**Figure 2**). You can open many files at the same time, which will pop up in a different window, and you will retain access to them through the recent files window every time you open JMP® (**Figure 2**). You have also the option of deleting files from the recent files window in case you do not need them accessible anymore.

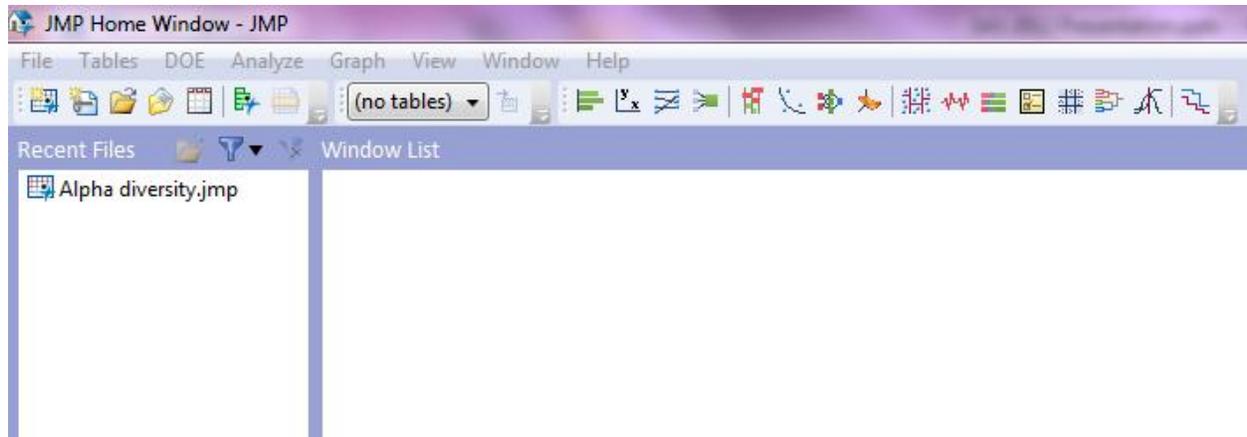


Figure 2 Top part of the JMP Home Window showing one file (Alpha diversity.jmp) in the “Recent Files” window.

If you click the file on the main JMP Home Window, the software will open the file in a different window (**Figure 3**). In this window, your data will show in a spread sheet format. On the left side of your data, you will find two sections: one for your columns and one for your rows. You can select columns directly from this panel (this is especially useful for large data sets) and if you select, exclude, hide, or label any row in your data sets, this information will be seen in the rows panel.

Graphical tools

There are many graphical tools that can be used in Microbial Ecology and other fields of study. Here, I will only talk about the Graph Builder, Chart, Overlay plot, Scatterplot 3D, and the Bubble plot.

Graph Builder. As the name implies, the Graph Builder is a tool that allows you to build your own graph. Although this may sound overly easy, there are several things that you need to consider before building your own graph. In this example, I will use the Alpha diversity.jmp file

to show how to use the Graph builder. First, choose and drag your (dependent) variable of interest into the y axis and your treatment column (column 1) into the x axis. Also, you can move Column 1 to the Color area to give each set of measurements (from the different treatment groups) a different color. To increase the marker size, simply right click on the central area of the graph, select graph and marker size. This is one thing that I do not like about the Graph Builder option. The biggest size you can choose is XXXL, which although may seem ok for the graph, sometimes I am in need of increasing the overall size of the graph for a publication, and if you do so, the marker size looks small. To increase the size of the axis' numbers or legends (in this example, the numbers for the Chao1 diversity index) and labels for the different treatments (control, Tx1 and Tx2), simply right click on the axis, choose axis settings and then font (here you can go up to a size 72). Now, here is the second thing I do not like in the Graph Builder. I just could not find a way to change the size of the labels of the axis (in this example, Chao1 for the y axis and Column 1 for the x axis). After a while, I discovered a way to do this. Just click Done (so you can exit the graph builder). Then select edit – all; then file – preferences, fonts – axis title. In here, you can change the size of the labels of the axis. I think this is a somewhat complicated way to do something that seems rather simple. If you find an easier way to do this, please let me know!

Chart. The Chart option within the Graph menu has also interesting characteristics. First, you will notice on the left lower corner an options panel, where you can choose bar, line, pie, needle or point charts, either vertical or horizontal. Simply choose the groups into the categories, X, levels. After you choose the type of chart you want, you can pick and move your dependent variables of interest into the “cast selected columns into roles” box. Interestingly, if you do not

pick what type of chart you want, you simply cannot move any column into the “cast selected columns into roles” box. Additionally, you have the option of selecting additional roles, which you may find useful for your own field of study. The charts you obtain are nice in size and colors (Figure 3), which you can change depending on your specific preferences.

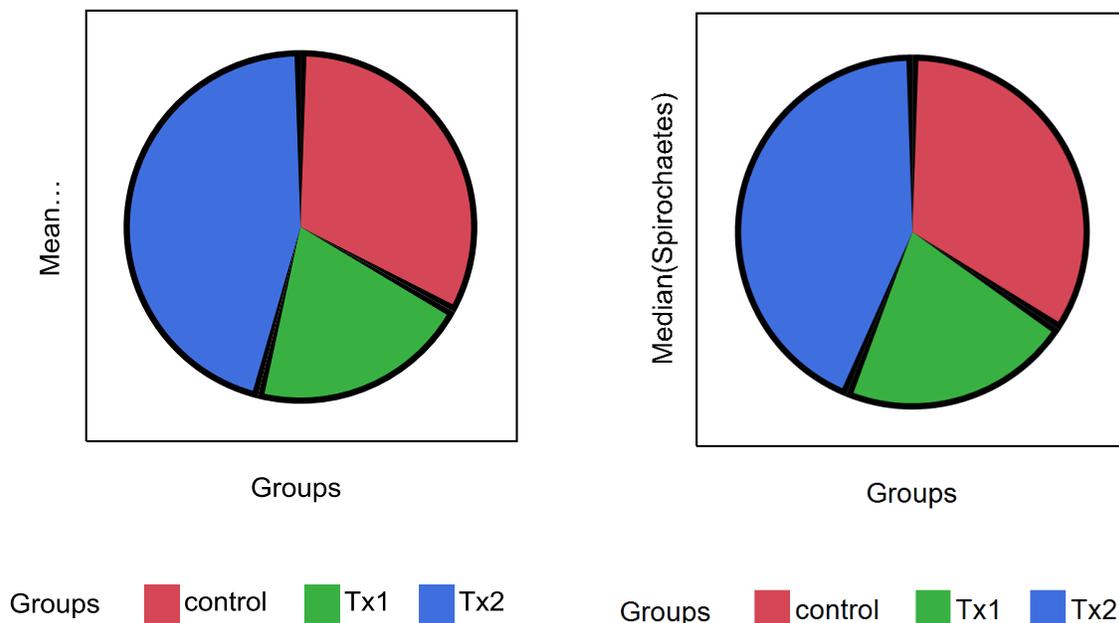


Figure 3 Charts obtained from the chart option in the graph menu. The pie charts show the mean (right) and the median (left) of a bacterial group (Spirochaetes) in the Control, Treatment 1 (Tx1) and Treatment 2 (Tx2) groups.

Overlay plot. The overlay plot is a useful graphical technique to visualize the relationship between two variables (x and y), with the options of including a number of additional independent factors. This is especially useful in the case you have a covariate (quantitative continuous independent variable) which may be linearly, quadratically or cubically related to your dependent continuous variable, a relationship that may additionally depend on additional independent variables (Figure 4 contains an example of this interesting case).

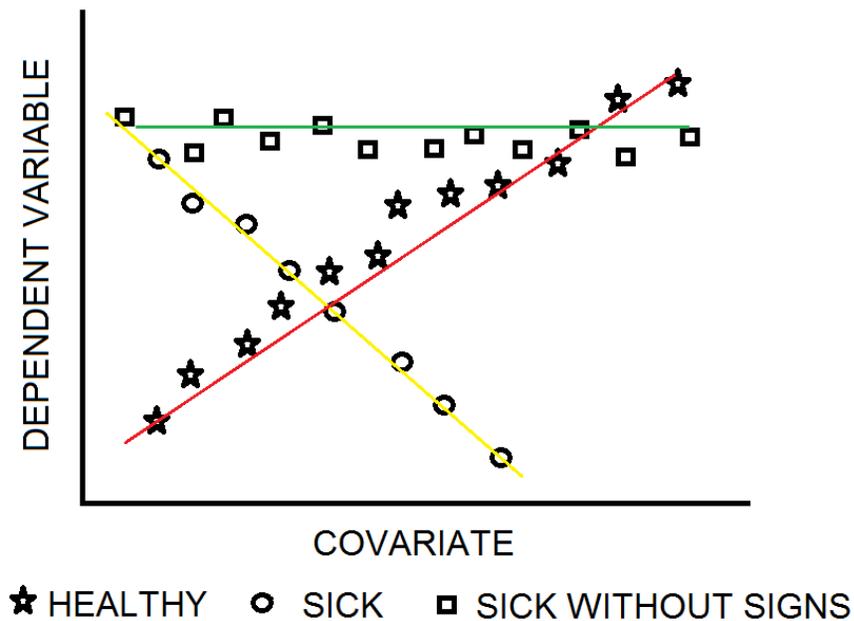


Figure 4 This figure shows a hypothetical relationship between a covariate (x axis) and a dependent variable (y axis). This relationship may be different among different levels of a third variable (in this case, health status: healthy, sick, and sick with no signs of disease).

Scatterplot 3D. I made a video about the scatterplot 3D that I will present in the SCSUG Educational Forum 2012. You can watch the video in YouTube using this link: <http://youtu.be/kvNrr5fS2ls>.

Bubbleplot

The Bubbleplot is perhaps my favorite graphical tool to look at the relationships between the abundance of different groups. For example, **Figure 5** allows you to look at the relationship of two separate bacterial groups (the Firmicutes and the Bacteroidetes). In addition, you can include a third group (in this case Verrucomicrobia). This I think is particularly useful in Microbial Ecology, because often we simply do not know if the relationship between two microbial groups

is affected by a third group. Other fields of study that I think the bubbleplot can come handy is sociology (one may look at the number of suicides in two culturally different areas of a country, sized by an assessment of religious preference) and genetics (the expression of a particular gene may be related to the expression of other gene, a relationship that may be depend on the expression of a third gene).

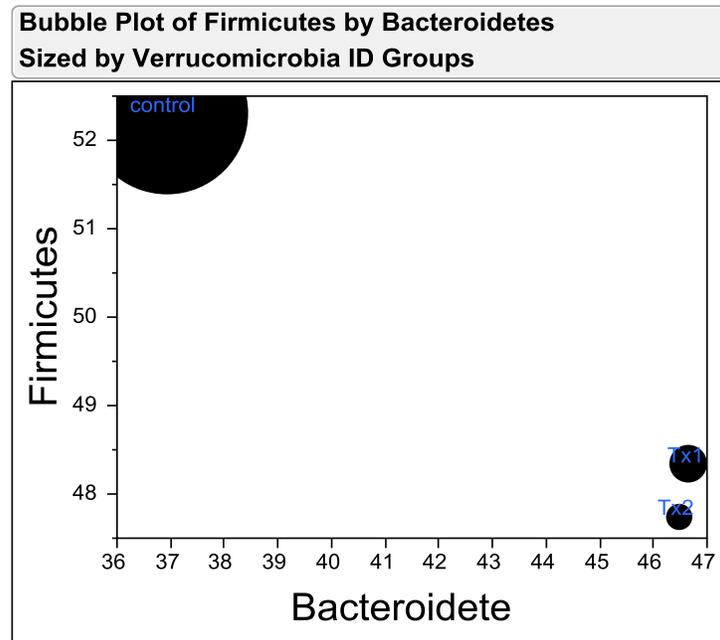


Figure 5 Bubble plot made in JMP showing the relationship between the abundance of two bacterial phyla (Bacteroidetes in the x axis, Firmicutes in the y axis), sized by the abundance of a third bacterial phyla (Verrucomicrobia).

Other graphical tool

There is another graphical tool that is not under the main Graph menu. Nevertheless, I consider this tool to be graphical in a sense that will show you a clustering of data.

Hierarchical clustering. This graphical tool is in the menu Analyze – Multivariate methods – Cluster. There are different methods to perform this type of clustering. For example, the Ward's minimum variance method clusters rows into groups (or clusters) whose values are close to each

other relative to those of other clusters. To illustrate the use of this method, let's use the sample data directory called the "birth data". Using this data set the Ward's minimum variance method clusters the values of birth rates that are more close to each other (Australia y Canada) (**Figure 6**). That is, compared to the rest of the countries, the birth rates of Australia y Canada are more alike. The birth rates values of Australia and Canada form then a cluster, and this process is repeated until there is only one cluster containing all the data points. It is important to note that JMP recommends hierarchical clustering for small data sets (a few hundred observations). I do not know whether JMP can manage more data (I do not see any reason why it should not), but I feel that the problem you could run into may be that many observations will create a clustering that is visually meaningless (i.e. you will not be able to distinguish who is who). The distance between two clusters in the different methods of hierarchical clustering does vary. For example, in the Ward's minimum variance method the distance between two clusters is the ANOVA sum of squares between the two clusters added up over all the variables, while in the average linkage method the distance between two clusters is the average distance between pairs of observations. More of this is available at the Help menu.

Statistical tools

Graphical tools are good to visualize the relationship or clustering of data points. However, the questions of whether there is a difference (for example among different treatments on intestinal microorganisms) or a relationship between different variables can only be solved by statistical methods. Here, I will revise two tools that are available in the Analyze menu: Fit Y by X, and Fit model.

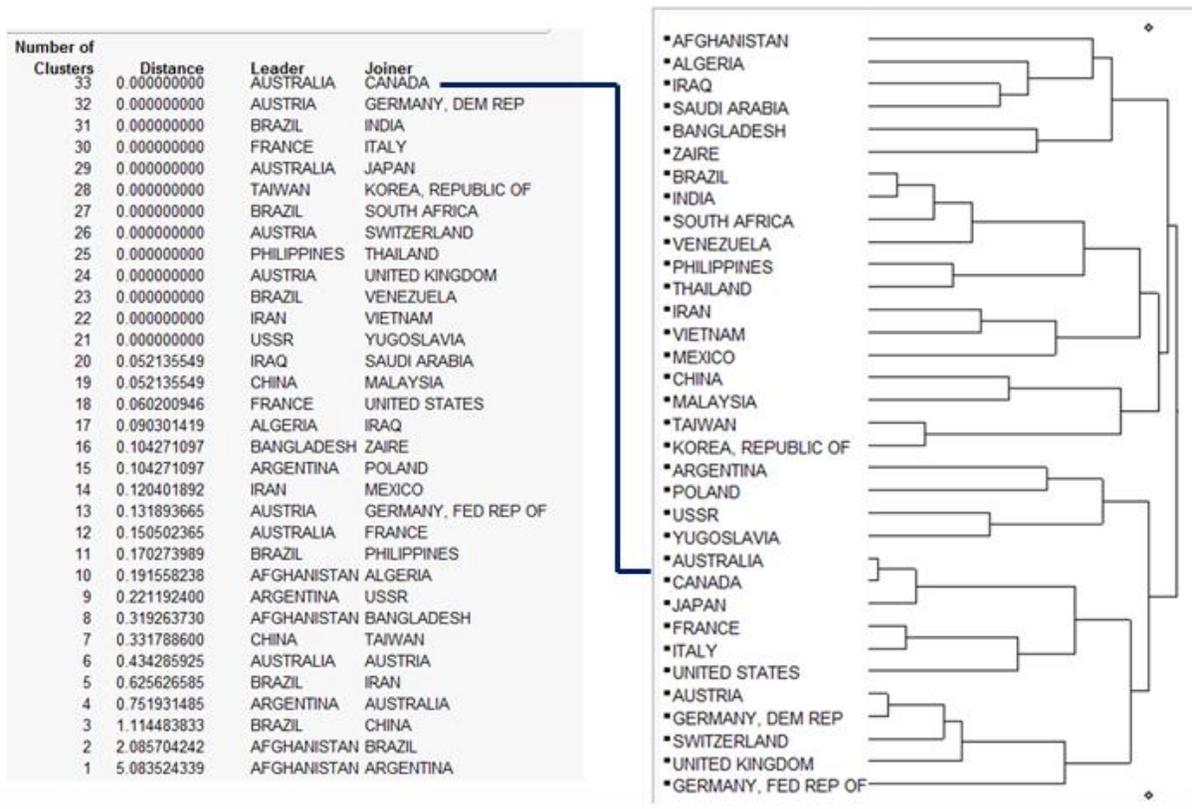


Figure 6 This figure shows a Hierarchical Cluster of Birth rates using the birth data available in the sample data directory of JMP®. The line is simply connecting the first cluster of values (Australia and Canada) to its place in the chart.

Fit Y by X. The Fit Y by X option is a quick tool to statistically analyze just as in a 1-way ANOVA. Either Y or X do not need to be continuous, and you can add a block factor. However, be aware that a blocking factor is normally treated as a random effect (e.g., in SAS) and the Fit Y by X option does not do that. If you have a blocking factor (e.g. cow in an experiment of feeding strategies to increase the quality of milk), you would need to use the Fit Model option (see below). A cool feature of the Fit Y by X is that when you choose your variables and place them in their boxes (Y, response; and X, factor), the software automatically decides what is the best way to analyze the data your data (Bivariate, Oneway, Logistic, or Contingency). I think this may be particularly useful for people that are not familiar with the different statistical procedures

but are nonetheless in need of a p-value to report. After you click Ok, you will see something like the **Figure 7**. An interesting feature of this results window is that you can choose a different type of analysis (for example, a non-parametric test) and different graphs to display your data.

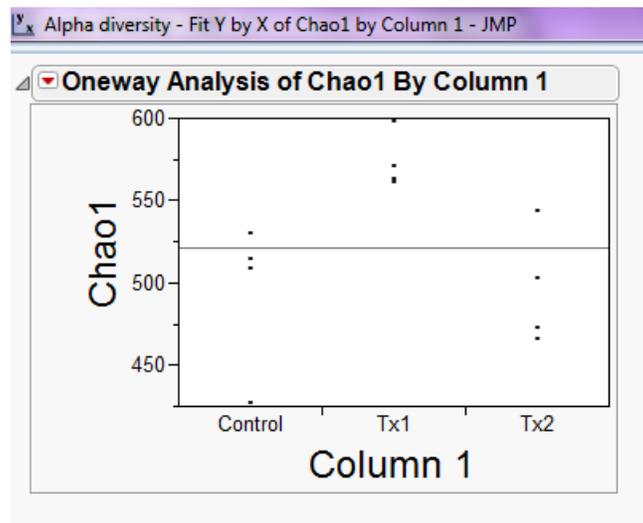


Figure 7 This figure shows a typical result from the X by Y option. The red triangle on the left top corner will display different graphical and statistical options.

Fit model. The Fit model option is more useful if you have a relatively large data set with different independent variables (continuous or discrete), potential interactions among them, and even better some random effects. I also discovered while doing this work that JMP offers generalized linear models (I've only done this in SAS), which is especially useful when your data is not normally distributed and you do not feel happy, like me, doing only nonparametric analyses. This Generalized Linear Model option is within the personality box along other options such as standard least squares, stepwise, manova and others. Within the generalized linear models you have many distributions such as normal, binomial and Poisson along several link functions such as identity, logit, probit, log, reciprocal, power. You can also assign attributes to your variables such as random effect, response surface effect or log variance effect. In a few

words, the Fit model option is an excellent option for experienced researchers having complex data sets. However, I highly recommend SAS to manage other things such as repeated measures and their associated covariance structures.

Summary

Microbial ecology is a fascinating topic as it encompasses the study of microorganisms in their natural environments. JMP® is a program that provides both graphical and statistical tools to analyze and learn from data. I hope this paper can be of some use for inexperienced users that have started using the software as well as for more experienced users in their journey through data discovery.

Listed References

- Duncan SH, Hold GL, Harmsen HJM, Stewart CS, Flint HJ (2002). Growth requirements and fermentation products of *Fusobacterium prausnitzii*, and a proposal to reclassify it as *Faecalibacterium prausnitzii* gen. nov., comb. nov. *International Journal of Systematic and Evolutionary Microbiology*, **52**: 2141-2146.
- Schulz HN, Jørgensen BB (2001). Big bacteria. *Annual Revisions in Microbiology*, **55**: 105-137.
- Sokol H, Seksik P, Furet JP, Firmesse O, Nion-Larmurier I, Beaugerie L, Cosnes J, Corthier G, Marteau P, Doré J (2009). Low counts of *Faecalibacterium prausnitzii* in colitis microbiota. *Inflammatory Bowel Disease*, **15**: 1183-1189.
- Whitman WB, Coleman DC, Wiebe WJ (1998). Prokaryotes: the unseen majority. *Proceedings of the Natural Academy of Science USA*, **95**: 6578-6583.

About the author

Jose grew up in Monterrey, the second largest city in Mexico. He studied Veterinary Medicine at the Universidad Autonoma de Nuevo Leon from 2000 to 2005, but since his third year he started helping out in the Microbiology Laboratory, fascinated by the growth of microorganisms in Petri dishes. In 2008, he had the honor to start a doctorate program in the Gastrointestinal Laboratory at Texas A&M University under the mentorship of one of the most enthusiasts gut Microbial Ecologists, Dr. Jan Suchodolski. Jose received the degree of Doctor of Philosophy in December 2011 in recognition for his work on the use of molecular techniques to study the gastrointestinal microbiota of cats and dogs. He has published 6 articles in international peer reviewed journals and more than 10 other scientific publications. He has also participated in many conferences, including the last three Educational Forums of the South Central SAS Users' Group and the last meeting of the International Scientific Association for Probiotics and Prebiotics. Jose currently serves as an Assistant Professor in the Faculty of Veterinary Medicine at the Universidad Autonoma de Nuevo Leon. He teaches a course on Biostatistics and Comprehension of Veterinary Scientific Texts in English. Jose is involved in a variety of different research projects about microorganisms in animals and can be contacted at: josegarcia_mex@hotmail.com. More information about Jose can be found at: sites.google.com/site/josefgarciamazcorro.