Data Analysis for Managers Handout

If history is any guide many students pick management as a major because there are "no numbers." This is unfortunate because numbers are a routine reality in the life of a manager. Being unfamiliar with the elementary statistics used in data analysis can easily result in poor decisions, missed opportunities, or you being taken advantage of by less ethical rivals. This handout is designed to refresh your understanding of statistics and to highlight things you may frequently see.

Charts & Graphs – Fun with Scales

If a picture is worth a thousand words then charts and graphs (preferably in full color and 3-D) are worth a thousand statistics! ☺ Any business presentation will be loaded with these. However, always check the axis, especially the y-axis, of the charts and graphs you are presented. Any change can be made to look dramatic by changing the axis/scale of the graphic used. By inverting the scale you can even make a decline seem like an increase. The Wall Street Journal modifies the axis all the time in their graphics. When looking at a graphic note the range of its y axis, if its bottom isn't zero the change being illustrated will look more dramatic. People can use this all the time to influence you and others.

Mean vs. Median – Manipulating Elementary Statistics

Almost everyone has a working knowledge of means because mean is just another name for an average – taking X numbers and then dividing them by X. People often get confused when confronted with statistics because statisticians think in math and write sentences in math instead of English. For example, in a statistics book mean is often expressed as:

$$X = (\sum x_{i....j})/n \text{ where } n= i....j$$

Intimidated? Don't be. All this says is add up all the numbers you have and then divide them by the number of numbers you added up. So the key thing here is to remember what we talked about concerning bluffing. If someone is using statistics and their meaning is not clear ask them to explain it to you in English. Anyone who actually knows what they are doing should be able to explain it to you without resorting to mathematical notation. Mathematical notation is just another way to express meaning (e.g. language) and can be translated into English just fine.

The median is frequently confused with the mean because they sound similar. Recall, the median is the exact middle number in a list of numbers when they are sorted. So if you had 1, 2, 3 the number 2 is the median. 2 would also be the mean in this case. But what if you had 1, 2, 97? 2 is still the median but 33 is the mean. This kind of skewing is common with lots of data.

Perhaps the most common instance of this occurs with national wealth data. Which do you think is higher in the United States, mean or median net worth? Believe it or not, according to the U.S. Census Bureau the mean net worth in the United States is dramatically higher than the median. For the year 2000 median net worth was $55,000 while the mean was a whopping

$182,381. The mean is so much higher because people like Bill Gates are REALLY rich while most of us are not quite so well off.

What's the point? Always ask for both of these numbers. Using only one in place of them together is a common way to "lie with statistics."[1]

Data Analysis – Simple Regression

The most common data analysis technique used in social sciences (of which management is a proud part!) is regression. It is very easy to understand and the statistical part is built right into EXCEL so you can use it at any time. Many other common statistical inference techniques can be subsumed by regression, so if someone is talking about ANOVA ask them to convert it into a regression analysis for you (if you are the big boss), it will be A LOT easier to understand.

Regression simply consists of fitting a line to a bunch of data. This is easy to picture in two dimensional or even three dimensional space. The great thing about regression is that mathematics is not constrained dimensionally so you can have lots of different variables that you can use to try and explain what's going on. The thing you're trying to explain is often called the dependant or experimental variable. The things you're using to try and explain the dependant variable are called independent or manipulation variables. Statisticians each have their own preferred terms for how they refer to these variables, so don't feel bad about making sure they are clear to you. I prefer dependant and independent.

When explained by statisticians regression is quite confusing. Statisticians usually try and explain regression mathematically, i.e. $(X'X)^{-1}X'Y$, using matrix notation. We're focused on data analysis so let's focus on how you can use regression. You use regression to see how one or more (independent) variables influence another single (dependant) variable. When you have just one (independent) variable you can just use a correlation matrix but when you have multiple (independent) variables regression is really easy to use.

To illustrate, let's look at some real data and a real issue. (See Handout) Given this data, can you see any relationships? Let's develop some theories about this data…

The data analysis that EXCEL[2] does is very good and looks overwhelming at first. The key is to focus on just a few critical items. I've listed them below in order of importance.

[1] There is actually a book with this title. Keep in mind that while it is possible, even easy, to lie with statistics it is A LOT easier to lie without them! You should feel free to contact me for additional books on this topic. A good one is Kennedy, A Guide to Econometrics.

[2] Excel has regression built into its Analysis Pack. It is menu driven and I can show you how to use it in less than five minutes if you are interested. Well, at least I can on Excel 98 for the Macintosh. ☺

First, find the **t Stat**. This tells you if a variable was statistically significant or not. Keep in mind you generally need 30 observations plus however many (independent) variables you have (_____ in this case) to get a good t test. Generally, any t statistic greater than 2[3] means that you can be 95% sure that the variable is statistically significant. The t statistic is simply the coefficient (see below) divided by the standard error. Some statisticians like to report standard errors instead of t statistics. In their tables this will appear in parenthesis underneath the coefficients. If they do not specify make sure you ASK if they are reporting t statistics or standard errors.

Second, find **Coefficients**. This tells you how much a one unit change in that independent variable influences the dependant one. However, remember that any variable that fails its t test is not significant, so while this is what your mathematical model says, don't take it to the bank.

Third, find **Adjusted R – squared**. This tells you how much of the relationship you are examining is explained by the variables you've picked. If you don't have any significant variables it's a good bet that this number is pretty low.

Finally, there is usually a **p value** somewhere in the output. This is the chance that the relationship you are examining is random chance.

Armed with these four statistics you should be able to make sense of most regression results. Keep in mind that these four numbers are likely to be buried beneath hundreds of other statistics and numbers in the output. Hence the importance of your eyes not glazing over when you see numbers because you've got to be alert to ask questions and look for the key things.

However, keep in mind there are many cautions:

1. Remember the 30 + number of variable rule for t tests. This is the problem with almost all medical research, they never have enough subjects to control for all relevant variables. This is why what is good for you changes from year to year.
2. Regression has A LOT of assumptions built into it. The most important ones are that the observations do in fact vary and that they are not correlated with each other. This is really common, so be careful. This happens a lot if you use prior period data as an independent variable, be careful.
3. Correlation is NOT causation. It is very easy to get the direction of the relationship wrong in regressions and all statistical inference for that matter. For example, I woke up today and sure enough the sun came up. The two variables are perfectly correlated. However, it would be crazy for me to suggest that I somehow "caused" the sun to rise. So be careful. This is why asking questions and not being intimidated by statistics is so important for managers and why I hate data mining, but that is a different story.

I hope you have found this summary beneficial. Keep in mind, people spend years training in statistics and frequently take multiple classes on just regression so do not assume expertise solely on the basis of this handout. The point is to explain some terms to help you have the confidence to ask questions when confronted with statistical "mumbo jumbo." Good Luck!

[3] Technically you need a probability distribution table to compute the exact number but 2 is a good guide.