Problems With Using Microsoft Excel for Statistics

Jonathan D. Cryer

(Jon-Cryer@uiowa.edu)

Department of Statistics and Actuarial Science

University of Iowa, Iowa City, Iowa

Joint Statistical Meetings

August 2001, Atlanta, GA

In this talk I will illustrate Excel's serious deficiencies in five areas of basic statistics:

Graphics

Help Screens

Computing Algorithms

Treatment of Missing Data

and

Regression

We begin with basic graphics.

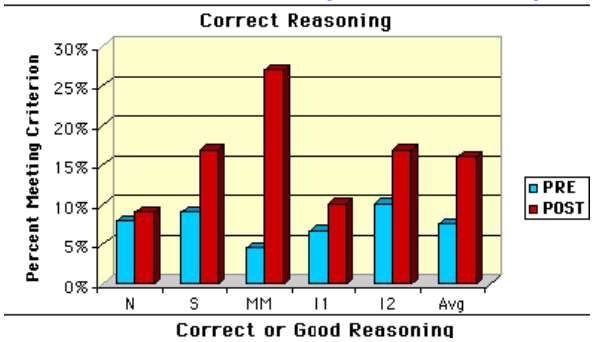
Good Graphs Should:

- ✔ **Portray Numerical Information Visually Without Distortion**
- ✔ **Contain No Distracting Elements (e.g., no** false **third dimensions nor "**chartjunk**")**
- ✔ **Label Axes (Scales) and Tick Marks Appropriately**
- ✔ **Have a Descriptive Title and/or Caption and Legend**

(References: Cleveland (1993, 1994) and Tufte (1983, 1990, 1997))

However, Excel meets virtually none of these criteria. As our first example illustrates, Excel offers false third dimensions on the vast majority of its graphs. (Unfortunately, this example is taken from the *Journal of Statistical Education*.)

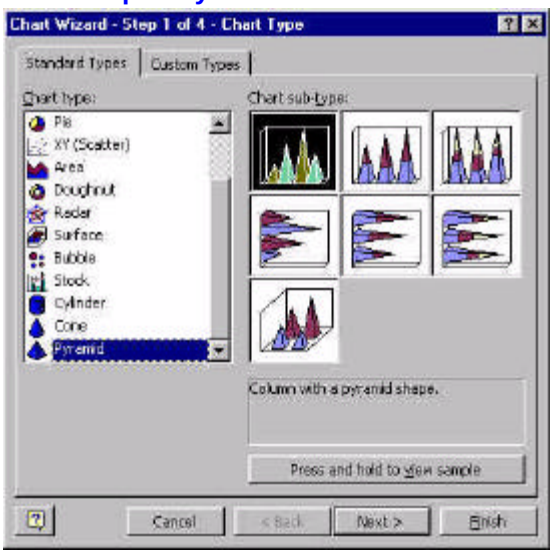
Example: Excel Graphics With False Third Dimension (taken from *JSE*!)



The vast majority of Chart types offered by Excel should NEVER be used!

Our next example shows the graph-types available as pyramid charts. None of these choices shown below represent good graphs! All but the last one display false third dimensions. In addition they all suggest stacked displays that are known to be poor ways to make comparisons.

Example: Pyramid Charts



(For the similar reasons, Excel's column, cone, and cylinder charts don't seem to have any redeeming features either!)

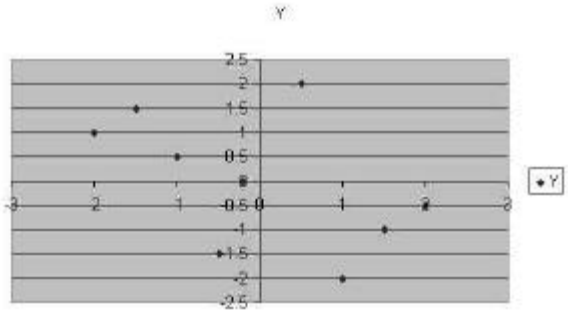
Scatterplots represent bread-and-butter graphs for visualizing relationships between variables.

Scatterplots Should Have:

- ✔ **Good Choice of Axes**
- ✔ **Meaningful Legends**
- ✔ **No False Third Dimensions**

However, Excel's default scatterplots leave much to be desired. In the following example two data points have been covered up by the axis labels. Can you find them? And is the legend displayed to the right of the graph useful? Note that there is no label for the horizontal axis.

Example: Excel Default Scatterplot



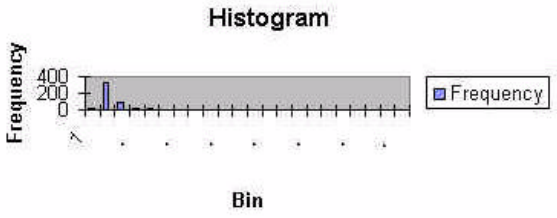
Histograms are another basic statistical display.

Histograms Should Have:

- ✔ **No Meaningless Gaps**
- ✔ **A Reasonable Choice of Bins**
- ✔ **An Easy Way To Choose Or Adjust The Bins**
- ✔ **A Good Aspect Ratio**
- ✔ **Meaningful Labels on Axes**
- ✔ **Appropriate Labels on Bin Tick Marks**

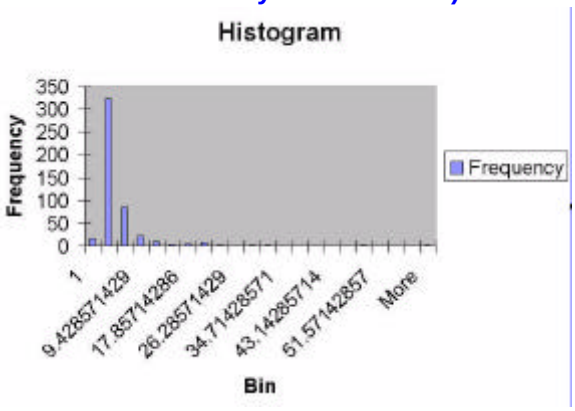
However, the next example shows a default histogram produced by Excel. The bin labels are impossible to read, the aspect ratio is poor, the legend and horizontal axis label are useless.

Example: Excel Default Histogram



If we click on the graph and stretch it vertically, we can then read the bin labels.

Example: Excel Histogram (stretched vertically to read labels)



The choice of class intervals or bins is rather bizarre, the number of digits displayed is atrocious, and it is not at all clear what tick marks these labels apply to.

In any software, the help screens should give useful and accurate information. In particular:

Help Screens Should:

- ✔ **Not Confuse**
- ✔ **Give Accurate Statistical Information**
- ✔ **Be Helpful!**

However, Excel's help for statistics is quite poor.

Here is an example of the Help screen displayed when you do a two-sample t-test.

Example: Excel 2000 Help Screen for the Two-sample T-Test

"t-Test: Two-Sample Assuming Equal Variances analysis tool

This analysis tool performs a two-sample student's t-test.

This t-test form assumes that the means of both data sets are equal; it is referred to as a homoscedastic t-test.

You can use t-tests to determine whether two sample means are equal."

These sentences contain a number of basic errors. About the only value in them would be to ask your students to critique them and locate the many errors!

The next example shows the help supplied for the confidence interval function.

Example: Excel 2000 Confidence Function

"CONFIDENCE

Returns the confidence interval for a population mean. The confidence interval is a range on either side of a sample mean. For example, if you order a product through the mail, *you can determine, with a particular level of confidence, the earliest and latest the product will arrive."* [emphasis mine]

The material emphasized, is, of course, a basic *misstatement* of the meaning of a confidence interval.

A last example displays the help given for the standard deviation function.

Example: Excel 2000 STDEV Function

"STDEV

Estimates standard deviation based on a sample. The standard deviation is a measure of how widely values are dispersed from the average value (the mean).

(snip...)

Remarks

(snip...)

The standard deviation is calculated using the "*nonbiased*" or "n-1" method.

STDEV uses the following formula:

$$\sqrt{\frac{n\sum x^2-\left(\sum x\right)^2}{n(n-1)}}$$

This help item introduces a new term, nonbiased, but that is the least of the difficulties here. (And, of course, the standard deviation given here is not unbiased for the population standard deviation under any set of assumptions that I know of!) More importantly, the formula given, the so-called "computing formula," is well-known to be a very poor way to compute a standard deviation. We return to this below.

Excel is especially deficient in its statistical analysis when some of the data are missing.

Treatment of Missing Data

✔ **Excel Does It Incorrectly**

✔ **Excel Does It Inconsistently**

✔ **Excel Makes Selecting Predictor Variables In Regression Especially Difficult When Data Missing**

As an example, here is a simple paired dataset with some of the data missing (NA= not available or missing):

Pre	Post
1	1
NA	2
3	3
4	NA
5	5
6	6
7	7
8	8
9	9

Here is the output of the paired data analysis of these data with the Excel Data Analysis Toolpack:

t-Test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	5.375	5.125
Variance	7.125	8.410714286
Observations	8	8
Pearson Correlation	1	
Hypothesized Mean Difference	0	
df	7	
t Stat	-1	
P(T<=t) one-tail	0.17530833	
t Critical one-tail	1.89457751	
P(T<=t) two-tail	0.35061666	
t Critical two-tail	2.36462256	

Means, variances, and df are all wrong (for paired data)! Nothing here is of much use! (But a naive user might not know or even notice!)

One of the well-documented deficiencies of Excel is its choice of computing algorithms.

Computing Algorithms for Basic Statistics

- ✔ **Excel Uses Poor Algorithms To Find The Standard Deviation (See Help screen for STDEV shown above)**
- ✔ **Excel Defines The First Quartile To Be The Ordered Observation At Position (n+3)/4**
- ✔ **Excel Does Not Treat Tied Observations Correctly When Ranking**
- ✔ **Regression Computations Are Often Erroneous Due To Poor Algorithms (See below)**

In addition Excel, usually displays many more digits than appropriate. (See histogram and paired t-test output shown above.)

Finally, Excel has major and documented difficulties with its regression procedures.

Regression in Excel

- ✔ **Does Not Treat Zero-Intercept Models Correctly**
- ✔ **Sometimes Gets Negative Sums Of Squares**
- ✔ **Does Not Handle Multicollinearity Correctly**
- ✔ **Computes Standardized Residuals Incorrectly!**
- ✔ **Displays Normal Probability Plots That Are Completely Wrong!**
- ✔ **Makes Variable Selection Very Difficult**

In summary:

Due to substantial deficiencies, Excel should not be used for statistical analysis. We should discourage students and practitioners from such use.

The following pretty much sums it up:

Get the Right Tool for the Job!

Friends Don't Let Friends Use Excel for Statistics!

References

Allen, I. E. (1999), "The Role of Excel for Statistical Analysis", Making Statistics More Effective in Schools of Business 14th Annual Conference Proceedings, ed. A. Rao, Wellesley: http://weatherhead.cwru.edu/msmesb/

Callaert, H. (1999), "Spreadsheets and Statistics: The Formulas and the Words", *Chance*, 12, 2, p. 64.

Cleveland, W. S., *Visualizing Data*, 1993, Hobart Press, Summit, NJ

Cleveland, W. S., *The Elements of Graphing Data*, Revised Edition, 1994, Hobart Press, Summit, NJ

Goldwater, Eva, Data Analysis Group, Academic Computing, University of Massachusetts, *Using Excel for Statistical Data Analysis: Successes and Cautions*, November 5, 1999, www-unix.oit.umass.edu/~evagold/excel.html

Knusel, Leo, “On the Accuracy of Statistical Distributions in Microsoft Excel 97”, *Computational Statistics and Data Analysis*, 1998, 26, pp. 375-377

McCullough, B.D. and Wilson B. (1999) "On the Accuracy of Statistical Procedures in Microsoft Excel 97", *Computational Statistics and Data Analysis*, 31, pp. 27-37.

McKenzie, Jr., J. D., and Rybolt, W. H. (1994), “What is the Most Appropriate Software for a Statistics Course?”, Computer Science and Statistics: Proceedings of Twenty-Sixth Symposium on the Interface, United States of America: Interface Foundation of North America.

_____ (1996), “Excel as a Statistical Package: Past, Present, and Future” presented at COMPSTAT '96, XII Symposium on Computational Statistics, Barcelona, Spain.

Sawitzki, Gunther, “Report on the Numerical Reliability of Data Analysis Systems”, *Computational Statistics and Data Analysis*, 1994, 18, pp. 289-301

Simon, Gary, ASSUME (Association of Statistics Specialists Using Microsoft Excel), untitled 19 page *Word* file, http://www.jiscmail.ac.uk/cgi-bin/wa.exe?A2=ind0012&L=assume&D=0&P=830

Simonoff, Jeffry, Stern School of Business, New York University, *Statistical Analysis Using Microsoft Excel*, 2000, www.stern.nyu.edu/~jsimonof/classes/1305/pdf/excelreg.pdf

Tufte, E. R., *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Conn., 1983

Tufte, E. R., *Envisioning Information*, Graphics Press, Cheshire, Conn., 1990

Tufte, E. R., *Visual Explanations*, Graphics Press, Cheshire, Conn., 1997