

Is Microsoft Excel an Adequate Statistics Package?

It depends on what you want to do, but for many tasks, the answer is 'No'.

Excel is available to many people as part of Microsoft Office. It contains some statistical functions in its basic installation. It also comes with statistical routines in the Data Analysis Toolpak, an add-in found separately on the Office CD. You must install the Toolpak from the CD in order to get these routines on the Tools menu. Once installed, these routines are at the bottom of the Tools menu, in the "Data Analysis" command. People use Excel as their everyday statistics software because they have already purchased it. Excel's limitations, and occasionally its errors, make this a problem. Below are some of the concerns with using Excel for statistics - Seven Deadly Sins - that are recorded in journals, on the web, and from personal experience.

There is more detail in addition to what's below on the Practical Stats website.

Limitations of Excel

1. Many statistical methods are not available in Excel.

This is Excel's biggest problem. Commonly-used statistics and methods NOT available within Excel include:

- Boxplots

- p-values for the correlation coefficient

- Spearman's and Kendall's rank correlation coefficients

- 2-way ANOVA with unequal sample sizes (unbalanced data)

- Multiple comparison tests (post-hoc tests following ANOVA)

- p-values for two-way ANOVA

- Levene's test for equal variance

- Nonparametric tests, including rank-sum and Kruskal-Wallis and more.

Excel's lack of functionality makes it difficult to use for more than computing summary statistics and simple univariate regression. Third-party add-ins to Excel attempt to compensate for these limitations, adding new functionality to the program (see "A Partial Solution", below).

2. Several Excel procedures are misleading.

Excel's regression routine provides a Normal Probability Plot option. However, it produces a probability plot of the Y variable, not of the residuals, as would be expected.

Excel's CONFIDENCE function computes z intervals using 1.96 for a 95% interval. This is valid only if the population variance is known, which is never true for experimental data.

Excel is inconsistent in the type of P-values it returns. Some are two-sided, some one-sided. Look carefully at the documentation of any Excel function you use, to be certain you are getting what you want.

For example, a one-sided t interval at $\alpha=0.05$, standard practice would be to look up the t-statistic in a textbook for $\alpha=0.05$. In Excel, the TINV function must be called using a value of $2*\alpha$, or 0.10, to get the value for $\alpha = 0.05$. Make sure you read the help menu descriptions carefully to know what each function produces.

3. Distributions are not computed with precision.

Statistical distributions used by Excel do not agree with better algorithms for those distributions at the third digit and beyond. So they are approximately correct, but not as exact as would be desired by an exacting statistician. This may not be harmful for hypothesis tests unless the third digit is of concern (a p-value of 0.056 versus 0.057). It is of most concern when constructing intervals (multiplying a std dev of 35 times 1.96 give 68.6; times 1.97 gives 69.0)

4. Routines for handling missing data were incorrect.

This was the largest error in Excel, but a 'band-aid' has been added in Office 2000. In earlier versions of Excel, computations and tests were flat out wrong when some of the data cells contained missing values, even for simple summary statistics. Error messages are now displayed in Excel 2000 when there are missing values, and no result is given. Although this is still inferior to computing correct results it is somewhat of an improvement.

5. Regression routines are incorrect for multicollinear data.

This affects multiple regression. A good statistics package will report errors due to correlations among the X variables. The Variance Inflation Factor (VIF) is one measure of collinearity. Excel does not compute collinearity measures, does not warn the user when collinearity is present, and reports parameter estimates that may be nonsensical. I find many examples of collinearity in environmental data sets, so doing multiple regression with Excel is dicey.

Excel also requires the X variables to be in contiguous columns in order to input them to the procedure. This can be done with cut and paste, but is certainly annoying if many multiple regression models are to be built.

6. Ranks of tied data are computed incorrectly.

When ranking data, standard practice is to assign tied ranks to tied observations. The value of these ranks should equal the median of the ranks that the observations would have had, if they had not been tied. For example, three observations tied at a value of 14 would have had the ranks of 7, 8 and 9 had they not been tied. Each of the three values

should be assigned the rank of 8, the median of 7, 8 and 9. Excel assigns the lowest of the three ranks to all three observations, giving each a rank of 7. This would result in problems if Excel computed rank-based tests. Perhaps it is fortunate none are available.

7. Many of Excel's charts violate standards of good graphics.

Use of perspective and glitz (donut charts?) violate basic principles of graphics. Excel's charts are more suitable to USA Today than to scientific reports. This bothers some people more than others.

A partial solution:

Some of these limitations (parts of 1,2,6 and 7) can be overcome by using a good set of add-in routines. One of the best is StatPlus, which comes with an excellent textbook, "Data Analysis with Microsoft Excel". With StatPlus, Excel becomes an adequate, though somewhat clunky, statistical tool. Without this add-in Excel is inadequate for anything beyond basic summary statistics and simple regression.

Data Analysis with Microsoft Excel by Berk and Carey
published by Duxbury (2000).

Opinion: Get this book if you're going to use Excel for statistics. (I have no connection with the authors of StatPlus and get no benefit from this recommendation. I'm just a satisfied user.)

References: *[URL links updated 9/2006]*

(1) On the accuracy of statistical procedures in Microsoft Excel '97

B.D. McCullough and B. Wilson, (1999), Computational Statistics & Data Analysis, 31,
pp 27-37

<http://www.elsevier.com/gej-ng/10/15/38/37/25/27/article.pdf>

or

<http://www.dia.fi.upm.es/~concha/excel.pdf>

(2) Problems with using Microsoft Excel for statistics [pdf Download]

J.D. Cryer, (2001), presented at the Joint Statistical Meetings, American Statistical Association, 2001, Atlanta Georgia

<http://www.cs.uiowa.edu/~jcryer/JSMTalk2001.pdf>

(3) Use of Excel for statistical analysis

Neil Cox, (2000), AgResearch Ruakura

<http://www.agresearch.cri.nz/Science/Statistics/exceluse1.htm>

(4) Using Excel for statistical data analysis

Eva Goldwater, (1999), Univ. of California – San Diego

<http://gcr.ucsd.edu/biostatistics/Excel.pdf>

(5) Statistical analysis using Microsoft Excel

Jeffrey Simonoff, (2002)

<http://www.stern.nyu.edu/~jsimonof/classes/1305/pdf/excelreg.pdf>

Guides to Excel on the web:

<http://www.rdg.ac.uk/ITS/Topic/Spreadsh/SpGExl9701/>

<http://www.rdg.ac.uk/ITS/Topic/Spreadsh/SpGExl9702/>

Practical Stats

<http://www.practicalstats.com>

-- Make sense of your data