Problems with using Microsoft Excel for Statistical Analysis & Graphics

Theresa A Scott, MS
Department of Biostatistics
theresa.scott@vanderbilt.edu
http://biostat.mc.vanderbilt.edu/TheresaScott

- CRC Research Skills Workshop -
Friday December 4, 2009
Microsoft Excel

- Widely available as part of Microsoft Office.
  - Automatically packaged with (ie, loaded on) new computers.
  - Version 2007 now available.

- Commonly used for data entry, storage, & management.

- Also commonly used for computation & graphics.
  - Basic installation contains a plethora of (built-in) functions, including some statistical functions.
    - Can also install a “Data Analysis Toolpak” add-in.
    - “Charting” capabilities also come with basic installation.

- These reasons encourage its use for statistical analysis.
Statistical Analysis in Excel

- “[I]t is quite possible that more basic statistical calculations are done worldwide in Excel than in all statistical packages combined”. – Wilkinson (1994)

- Unfortunate, since Excel is **not** a statistical package.

- Has been known for a long time (in the small world of statistical computing) that serious errors exist in Excel’s statistical procedures.
  - Since 1994 – when Sawitzki published two manuscripts in *Computational Statistics and Data Analysis*.

- Remained Microsoft’s dark secret in the larger world.
Statistical Analysis in Excel, cont’d

- Many textbooks with titles like “Statistics with Excel” have been written by professional statisticians.

- Many professional statisticians (continue to) use Excel on a daily basis for quick and easy statistical calculations.

- In turn, a generation of students learned to do statistics with Excel.
  - “Surely”, the student reasoned, “it is safe to use Excel for statistics. If it weren’t, my professor would have chosen a different software package.”
  - Subsequently went on to use Excel in the professional world.

- Big question: Has Microsoft addressed/fixed these errors?
Microsoft’s track record

- Typically, when errors are found in a software package (like SAS or Stata), the developer fixes the errors, & the matter is done.
  - No further need to re-evaluate the software on this set of tests.

- Not so with Microsoft & Excel.
  - Occasionally fixes errors, more often ignores them, & sometimes fixes them incorrectly.
    - Flaws identified by Sawitzki in 1994 have never been fixed.
    - List of errors of which Microsoft is aware is not available.
  - Consequently, every time there is a new version of Excel, the tests must be repeated – “Is it safe to use Excel?”
  - Also, not all statistical flaws have been found – only a fraction of the statistical functions & procedures have been tested.
Individual Problems

- NOTE: The following is *not* a comprehensive list.
Numerical Accuracy

- **Problem (from Excel 97):** The algorithms Excel uses to calculate statistical distributions do not agree with better algorithms for those same distributions at the third decimal place & beyond.
  - So, p-values are approximately correct, but not as exact.
  - Harmful for hypothesis tests if third decimal place is of concern (eg, a p-value of 0.056 vs 0.057).
  - Is of most concern when constructing confidence intervals (eg, $35 \times 1.96 = 68.6$ compared to $35 \times 1.97 = 69.0$).

- Example of other general problem that affects (at least some versions of) Excel 2007:
  - Formula “=$850 \times 77.1$” returns the value 100,000 not the correct value of 65,535 (but sometimes it does)!
Some Descriptive Statistics

• Problem (with Excel 2003 & older versions): The STDEV (standard deviation; SD) & VAR (variance) functions employed an internal formula that calculated the wrong result under specific situations – for variables whose mean is large compared to their SD.

• Affected calculation of other descriptive statistics (using other statistical functions).

• Seems OK in Excel 2007.

<table>
<thead>
<tr>
<th>Observation</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10,000,000,001</td>
</tr>
<tr>
<td>2</td>
<td>10,000,000,002</td>
</tr>
<tr>
<td>3</td>
<td>10,000,000,003</td>
</tr>
<tr>
<td>4</td>
<td>10,000,000,004</td>
</tr>
<tr>
<td>5</td>
<td>10,000,000,005</td>
</tr>
<tr>
<td>6</td>
<td>10,000,000,006</td>
</tr>
<tr>
<td>7</td>
<td>10,000,000,007</td>
</tr>
<tr>
<td>8</td>
<td>10,000,000,008</td>
</tr>
<tr>
<td>9</td>
<td>10,000,000,009</td>
</tr>
<tr>
<td>10</td>
<td>10,000,000,010</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>10,000,000,005.5</td>
</tr>
</tbody>
</table>

STDEV (in old Excel) 0.000

Correct STDEV 3.028
“Problem” (with all versions): Difference in the way Excel calculates percentiles from the way other statistical packages (may) calculate them.

• Using the QUARTILE function in Excel, 1\textsuperscript{st} & 3\textsuperscript{rd} quartile = 130 & 157.5.
• Using another stat package (SPSS), 1\textsuperscript{st} & 3\textsuperscript{rd} quartile = 125 & 160.
• Discrepancies have to do with how Excel calculates ranks – does not take into account “tied” ranks.
• Thus, also discrepancies between calculated value of the median (50\textsuperscript{th} percentile).

<table>
<thead>
<tr>
<th>Observation</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>120</td>
</tr>
<tr>
<td>2</td>
<td>125</td>
</tr>
<tr>
<td>3</td>
<td>125</td>
</tr>
<tr>
<td>4</td>
<td>145</td>
</tr>
<tr>
<td>5</td>
<td>145</td>
</tr>
<tr>
<td>6</td>
<td>150</td>
</tr>
<tr>
<td>7</td>
<td>150</td>
</tr>
<tr>
<td>8</td>
<td>160</td>
</tr>
<tr>
<td>9</td>
<td>170</td>
</tr>
<tr>
<td>10</td>
<td>175</td>
</tr>
</tbody>
</table>
Missing Data

- **Problem:** Missing values handled inconsistently & incorrectly.
  - In older versions, any blank cell considered to be zero.
  - At some point, changed so that blank cells were ignored (but still not in all cases).
  - However, paired t-tests, ANOVA, Regression & other Data Analysis Toolpak tools in badly deal with missing values.
  - Seems not be the case for built-in statistical functions.

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

(Correct) p-value from TTEST (PAIRED) 0.0448
p-value from Toolpak 0.2044
Other Misleading/incorrect Methods

- CONFIDENCE function (even in Excel 2007) uses incorrect constant in calculating a 95% confidence interval (CI).
  - General formula for 95% CI is mean ± constant*SD.
  - Excel uses 1.96 as the constant (ie, assumes a normal distribution).
  - Valid only if the population variance is known - never true for experimental data.
  - Thus, CIs computed on sample data will be too small.
    - Constant based on t-distribution should be used instead.

- Excel is inconsistent in the type of p-values it returns (by default).
  - Most often returns one-sided p-values.
  - However, in the TINV function, Excel returns a 2-sided p-value.
Other Misleading/incorrect Methods

- Bad linear regression algorithm(s).
  - Until Excel 2003, Excel would give completely incorrect coefficient estimates.
    - Algorithm was improved in Excel 2003 – both Excel 2003 & 2007 report correct coefficients.

- However, regression routines are still incorrect for *multicollinear data* – thus, (still) return incorrect coefficient estimates.
  - A good statistics package will report errors due to correlations among the X variables (covariates; predictors).
  - Excel does not compute any collinearity measures, & does not warn the user when collinearity is present, but does report (often nonsensical) coefficient estimates.
Excel’s Charts

- **Problem:**
  - Default chart types in Excel 2007 (and older) violate standards of good (statistical) graphics (defined by Tufte, Cleveland, etc).
  - Instead, these charts create *chartjunk* – hinder your ability to comprehend the data (i.e., extraneous graphical elements).
  - **Examples:** pie & donut charts; area charts; bubble charts; cylinder, cone, & pyramid bar & column charts; & any “3-D” charts.

- **Solution:**
  - Get to know the principles of good graphing well enough so you know how to choose appropriate options to override defaults.”
  - “Just because you can doesn’t mean you should.”
Excel’s Charts – *Examples*

Pie chart

Area chart

Line chart

Bubble chart
Other General Remarks

- Many statistical methods (& graphs) simply not available in Excel.
  - Histograms & boxplots;
  - test of significance for a correlation coefficient;
  - Spearman’s & Kendall’s rank correlation coefficients;
  - non-parametric tests of association (eg, Wilcoxon rank-sum & Kruskal-Wallis);
  - 2-way ANOVA with unequal sample sizes (ie, unbalanced data);
  - GLM (generalized linear models);
  - Survival analysis methods; &
  - regression diagnostics.

- Makes it difficult to use it for more than computing summary statistics, simple linear regression, & (some) hypothesis testing.
Other General Remarks, cont’d

- Inconsistencies between Data Analysis Toolpak add-in & built-in statistical functions, including *many more* errors unique to Toolpak.

- Some general issues specific to Toolpak:
  - Data organization differs according to analysis – must reorganize your data in many ways.
    - Example: Require the X variables (ie, predictors; covariates) to be in contiguous columns in order to input them to the regression procedure.
  - Many analyses can only be done one column at a time, making it inconvenient to do the same analysis on many columns.
  - Output is poorly organized, sometimes inadequately labeled, & there is no record of how an analysis was accomplished.
Many reported (& some allegedly fixed) bugs necessarily provoke suspicion that others still exist or may have been introduced.

No information about the nature of the numerical algorithms employed is generally provided or can be found.
- Help-files provided, but are often confusing, provide inaccurate statistical information, and/or are not helpful.

Excel does not provide any record (ie, log or history) of what is done (including any changes to your data), making it virtually impossible to document or duplicate what is done.
- Vital for serious analysis & key to reproducible research.
Conclusions & Recommendations

- Excel is a poor choice for statistical analyses beyond the simplest descriptive statistics, or for more than a very few columns.
  - Don’t assume that Excel will give the correct answer.
  - The “Data Analysis Toolpak” is not worth bothering with either.
    - No easier to use than most statistics package, has very limited capability, & also has (many) known bugs.

- Use a *real* statistical package when you need to do statistics.
  - SPSS, Stata, or SAS (see Vandy ITS for price), or R (free).

- Spreadsheet alternatives: *OpenOffice Calc & Gnumeric* (both free).
  - Both read-in Excel spreadsheets (ie, .xls files).
  - Some Excel errors not an issue, but no formal testing done yet.
Other Stats Resources at Vandy

- Biostatistics Clinics
  - Daily from 12:00 – 1:15 PM.
  - See http://biostat.mc.vanderbilt.edu/Clinics for more info.

- VICTR Resource Request
  - VICTR = Vanderbilt Institute for Clinical & Translational Research.
    - Supported by the Vanderbilt Office of Research & an NIH sponsored Clinical & Translational Science (CTSA) Award.
  - One option, a “Voucher”, grants you $2000, which can be used to work with a Biostatistician (covers ~20 hours of their time).
    - Before your data is collected and/or once your data is collected.
  - See the “Funding Support” tab once logging into StarBrite webpage (http://www.mc.vanderbilt.edu/starbrite) for more info & to submit a request (usually approved within 2-3 business days of submission).
References

References, cont’d

- *Should Microsoft Excel Software Be Used for Statistical Analysis or Graphics?* (http://andrologi-indonesia-pandi.org/_UPLOAD_/article_43817_Excel.pdf)

- http://biostat.mc.vanderbilt.edu/ExcelProblems