# Software for Intro Stats: Is Excel an Option?

Roger L. Berger
Arizona State University

August 2006

New Researchers Conference

Seattle, WA

# OUTLINE

1. Course description
2. Survey
3. Why use Excel?
4. Why not use Excel?
5. Conclusions
6. Bonus survey
7. Other topics

## Course Description

- Undergraduate, *The Basic Practice of Statistics*, David Moore
- Students from many different disciplines
- Maybe fulfills a general studies requirement
- Teaches methods
  - ⋈ descriptive statistics
  - ⋈ populations & samples
  - ⋈ $t$ tests & confidence intervals
  - ⋈ oneway ANOVA
  - ⋈ simple linear regression
  - ⋈ $\chi^2$ tests for contingency tables
- Not a “concepts” course

# Survey

*Choose one answer.* What kind of software/technology should be used in this kind of course?

A. None, use only pencil & paper

B. A hand calculator, nothing more

C. Educational software that comes with the text, e.g., CrunchIt

D. Microsoft Excel

E. Full-featured statistics package, e.g., SAS, SPSS

F. Statistics programming language, e.g., R, S-Plus

## Some Factors to Consider

1. Accessible – easily available to students? school? home? during exams?
2. Easy to use
3. Serviceable – do what is needed for this course?
4. Affordable
5. Useable after this course?

# Why Use Excel?

1. Most students have access to Excel on school computers
2. Many students have access to Excel on personal computers
3. No additional expense to students
4. Many students have used Excel to some degree
5. Excel has built-in tools that will carry out most of the analyses that are covered in this course
6. Many resources related to Excel; textbooks, online, etc.
7. Familiarity with Excel is desired by many employers

Hot off the press:

Nash, J. C. (2006). Spreadsheets in Statistical Practice–Another Look, *The American Statistician*, **60**, 287-289.

Spreadsheets are ubiquitous. We should work to ensure that they are used correctly for statistical practice.

Statistical analysis methods are provided in two forms in Excel:

1. Functions – give a single number as output, e.g.,
   - `=average(a1:a30)`
   - `=stdev(a1:a30)`
   - `=percentile(a1:a30,.6)`
2. Procedures – in the Data Analysis Toolpack add-in
   - two-sample $t$ tests
   - ANOVA – oneway and twoway
   - multiple regression

## Why Not Use Excel?

1. Poor graphics
   - ⋈ no boxplot or stemplot (many online, e.g., boxplot by Dawson from Moore & McCabe webpage)
   - ⋈ poor and inaccurate histograms
2. Some missing methods
   - ⋈ no one sample $t$ tests or confidence intervals
   - ⋈ no $\chi^2$ tests for contingency tables
3. Poor descriptions of some methods
   - ⋈ ZTEST function

4. Computational inaccuracies

   - ⋈ use of poor algorithms in earlier versions
   - ⋈ worst problems corrected in 2003 version
   - ⋈ accuracy probably adequate for this course

5. Limited capabilities beyond this course. Cannot do serious data analysis.

   - ⋈ no 3-way ANOVA
   - ⋈ no logistic regression
   - ⋈ rigid input formats, e.g., multiple rgression

Before the 2003 version, Excel used the “hand calculation” formula for calculation of the sample variance

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}$$

This formula leads to computational inaccuracies if $n$ is large and the $x_i$’s cover a wide range.

## References for Problems with Statistical Computations in Excel

McCullough, B. D., and Wilson, Berry (1999). On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics and Data Analysis*, **31**, 27–37.

. . . and later papers by these authors in 2002 and 2005

Website by David A. Heiser
www.daheiser.info/excel/frontpage.html
Microsoft Excel 2000 and 2003 Faults, Problems, Workarounds and Fixes (Detailed and up-to-date descriptions.)

## Conclusions – Some Positives

- Reasonable computational tool for an introductory, undergraduate, statistical methods course
- Accessible to most students, even at home
- Performs most computations needed for intro course
- Easy to introduce because many students are familiar with Excel
- Excel familiarity useful in other contexts

## Conclusions – Some Negatives

- Not appropriate for advanced data analysis
- Not appropriate for undergrad statistics majors
- Not appropriate for graduate students
- Some accuracy issues

# Bonus Survey

Evaluate the following expressions:

1. $-3 * 3$
2. $0 - 3 * 3$
3. $0 - 3^2$
4. $-3^2$

# Bonus Survey

Evaluate the following expressions:

| | | Answers |
|---|---|---|
| 1. | $-3 * 3$ | $\mathbf{-9}$ |
| 2. | $0 - 3 * 3$ | $\mathbf{-9}$ |
| 3. | $0 - 3^2$ | $\mathbf{-9}$ |
| 4. | $-3^2$ | $\mathbf{+9}$ |

What's going on?

- In Excel, *unary negation*, as in $-3^2$, has a higher order of precedence than exponentiation or any other operation
- $-3^2$ is interpreted as $(-3)^2$

Two high school students, Donald Brandl & Jebina Rajbhandari, discovered this when considering an example from McCullough & Wilson (1999).

They were trying to use the Excel Solver Toolbox add-in to fit this nonlinear least squares regression.

$$y = \beta_1 e^{-\beta_2 x} + \beta_3 e^{-(x-\beta_4)^2/\beta_5^2} + \beta_6 e^{-(x-\beta_7)^2/\beta_8^2} + \epsilon$$

About Excel: Questions? Comments?

Other topics?