Simple Statistics with Excel

Aniko Szabo, PhD
Associate Professor of Biostatistics

Made possible by the

Clinical and Translational Science Institute (CTSI),

and the

Division of Biostatistics

Speaker Disclosure

In accordance with the ACCME policy on speaker disclosure, the speaker and planners who are in a position to control the educational activity of this program were asked to disclose all relevant financial relationships with any commercial interest to the audience. The speaker and program planners have no relationships to disclose.

CME Evaluations!

Please help us by filling out an evaluation even if you are not eligible for CME credit.

Getting data into Excel

- Data can be
  - entered directly into Excel
  - imported from an existing file (text, Access)
  - imported from a Web-page
  - copy-pasted from Word, Acrobat, etc
    - if Excel puts it into one column, use the Text-to-Columns Wizard
- Many of these features are accessed through the Data tab

Data structure

- All data should be structured as a **list**:
  - each cell contains one value
  - each column contains one variable
  - the physical arrangement, spacing, color, etc should not carry additional information
  - each row contains information on one subject
  - each row is self-contained
- Do not mix data with analyses
- Missing values should be empty cells

Converted to a list

- Each row is one experimental unit
- Group is repeated for every subject
- Variable names have no special characters
- Averages/standard deviations are not part of the data
- Missing values coded consistently
- E

Transforming data

- Use Excel formulas for calculations
  - any cell that starts with an "=" sign is interpreted as a formula
- Create a new column for the transformed value
- Refer to values by column/row: A3, B12
- Refer to ranges as topleft:bottomright A2:C4

	A	B	C	Contents of column C
1	Cholesterol	HDL	Ratio	
2	291	30	9.7	=A2/B2
3	209	27	7.7	=A3/B3
4	272	27	10.1	=A4/B4

formula auto-updates if copied

Descriptive statistics

- Built-in functions can be used:
  - AVERAGE
  - MEDIAN
  - STDEV
  - “Insert function” on Formulas tab

	A	B
1		Group 1
2		Cholesterol (mg/dL)
3		291
4		209
5		272
6		293
7		302
8		304
9		
10		
11	AVG	278.50
12	SD(n−1)	35.9
13	SEM	14.7
14	N	6

=AVERAGE(B3:B8)
=STDEV(B3:B8)
=B12/SQRT(B14)
=COUNT(B3:B8)

Cross-tabulations

- Pivot tables give very good one- or multi-way tables
  - Can show frequencies, but also means, sums of one variable grouped by other variables
  - Found on “Insert” tab
  - Will update when “Refresh” is pressed

	Values					
Row Labels	Count of Group	Percent in group	Average of Cholesterol	StdDev of Cholesterol	Average of HDL	Average HDL as percent of group 1
1	6	20.0%	278.5	35.9	26.3	100.0%
2	8	26.7%	290.1	74.1	20.5	77.8%
3	7	23.3%	269.0	41.7	27.2	101.4%
4	9	30.0%	254.7	33.9	27.5	101.7%

Correlations

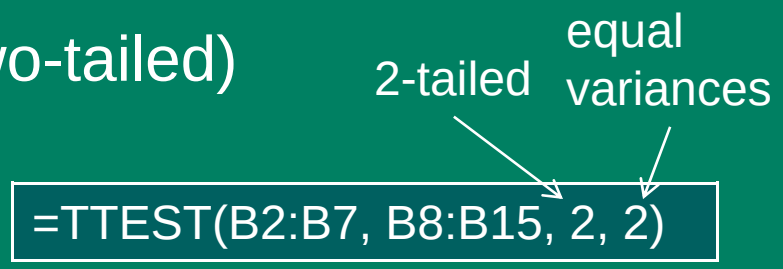
- Data Analysis Tool > Correlations
  - gives matrix of Pearson's correlation coefficient for a contiguous set of columns
  - no sample sizes, p-values
- CORREL function
  - will calculate correlation coefficient for any two columns
- Cannot compute Spearman correlation
  - The ranking tool and the RANK function give incorrect results for tied values

Statistical inference

- Essentially no support for categorical data analysis (confidence interval for proportion, chi-square test, etc)
- t-test, ANOVA, regression are available through Data Analysis Tools and/or functions
  - function can behave differently from add-in
  - have numerical instabilities, and should not be used for large problems

Two-sample t-test

- Requires values for each group to be contiguous
  - Data might have to be sorted
  - Excel tries to ensure that entire data row is sorted
- Data Analysis Tools > t-test: two-sample assuming equal/unequal variances
- TTEST function
  - gives only p-value (one- or two-tailed)
  - TYPE=2: equal variances
  - TYPE=3: unequal variances
- No confidence interval



Paired t-test

- Data Analysis Tools > t-test: paired samples
- TTEST function
  - gives only p-value (one- or two-tailed)
  - TYPE=1: paired
- Missing values are handled incorrectly by the Data Analysis Tool (but not the TTEST function)

Linear regression

- Data Analysis Tools > Regression
  - predictors (x variables) have to be in contiguous columns
  - no missing values allowed
  - do NOT run a regression through the origin
- In a scatter plot a trend line can be added with equation shown

Limitations of Excel

- Potential problems with analyses involving missing data
- Varying expectations regarding the arrangement of data
- Output scattered in many different worksheets, or all over one worksheet
- Output may be incomplete or may not be properly labeled
- No record of what you did to generate your results

Right tool for the right job

Excel is not the right tool for all but the simplest analyses

Resources

- The Clinical and Translation Science Institute (CTSI) supports education, collaboration, and research in clinical and translational science: www.ctsi.mcw.edu
- The Biostatistics Consulting Service provides comprehensive statistical support www.mcw.edu/biostatistics.htm
-

Free Drop-In Consulting

- MCW/Froedtert/CHW:
  - Monday, Wednesday, Friday 1 – 3 PM @ Froedtert Pavilion-TRU (LL772A)
  - Tuesday, Thursday 1 – 3 PM @ Health Research Center, H2400
- VA: Monday, 9:30-10:30 am
  - VA Medical Center, Room 70-A 314-A
- Marquette: Tuesday, 8:30-10:30 am
  - School of Nursing, Clark Hall, Office of Research & Scholarship