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

Institute *for*
Health and Society

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.

Outline

- Data entry
- Descriptive statistics
  - means
  - cross-tabulation
- Statistical inference
  - t-test
  - regression

Note: all specifics are for Excel 2007

Data Analysis Tools

- Many statistical analyses are available through the Data Analysis Add-in
- To install:
  - Office Button
  - Excel Options (button at bottom right)
  - Add-Ins tab
  - at bottom: Manage Add-Ins > Go…
- Will appear on the Data tab

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
- Extra notations removed

1	291	30
1	209	27
1	272	27
1	293	27
1	302	24
1	304	23
2	386	19
2	208	24
2	250	21
2	246	24
2	214	29
2	292	20
2	326	9
2	399	18
3	311	22
3	248	37
3	279	n/a
3	256	22
3	215	26
3	334	26
3	240	30
4	249	21
4	252	26
4	237	27
4	231	23
4	311	24
4	197	n/a
4	269	33
4	252	33
4	294	33

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 tr24 375tf 27.984hed1 s

Descriptive statistics

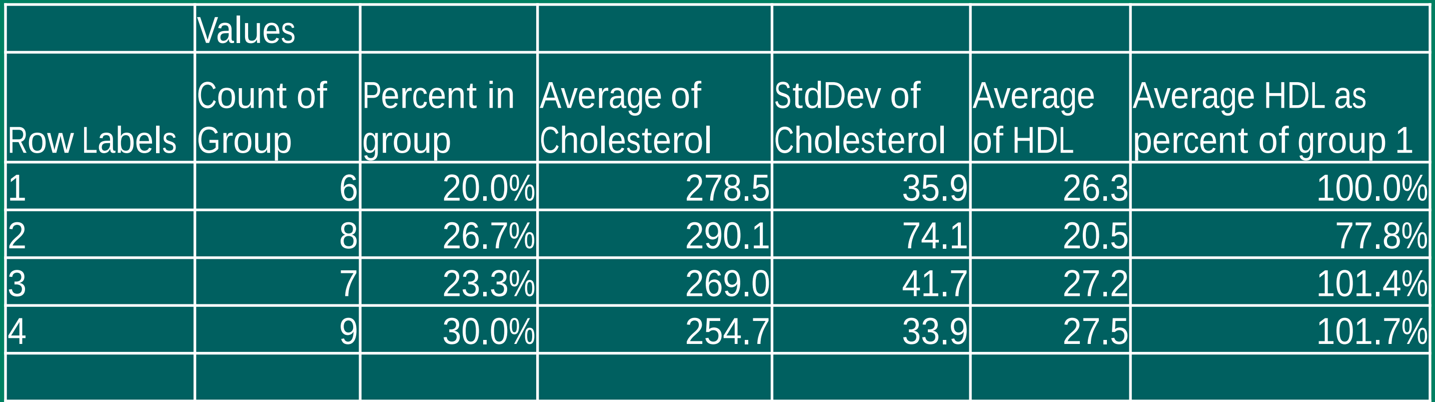
- Built-in functions can be used:
  - – AVERAGE
  - – MEDIAN
  - – STDEV
  - – “Insert function” on Formulas tab

Descriptive statistics

- In Data Analysis Tools: Descriptive statistics
  - – don’t put the result on the same page

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

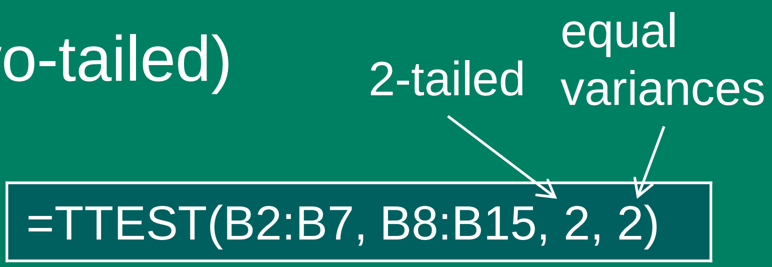
- Data Analysis Tool > Correlations
  - gives matrix of Pearson’s correlation coefficient for a contiguous set of columns
  - no sample sizes, p-values
- CORREL function
  - will calculate correlation coefficient for any two columns
- Cannot calcu

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



2-tailed

equal variances

=TTEST(B2:B7, B8:B15, 2, 2)

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)

ANOVA

- Balanced one- or two-way ANOVA available in Data Analysis Tools, but requires different data arrangement
- Unbalanced (or balanced) ANOVA can be run using the regression module: instead of Group, use G2, G3, and G4 as predictors

	A	B	C (formula)	C	D (formula)	D	E (formula)	E
1	**Group**	**Y**		**G2**		**G3**		**G4**
2	1	291	*=IF(A2=2,1,0)*	0	*=IF(A2=3,1,0)*	0	*=IF(A2=4,1,0)*	0
3	2	386	*=IF(A3=2,1,0)*	1	*=IF(A3=3,1,0)*	0	*=IF(A3=4,1,0)*	0
4	3	311	…	0		1		0
5	4	249		0		0		1
6	…	…						

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

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