

Introduction to Excel for Statistical Analysis

Taught by Sean P. Rogers & Sarah Morrell
MGSC 2301: Business Statistics
Professor Sahar Abi-Hassan
Fall 2024



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Workshop Agenda

Learning Objectives

- Understand relevant uses of Excel
- Learn about basic Excel functions and vocabulary
- Making tables and chart
- R Demonstration

Slides, data, and code available at the link below:

[https://bit.ly/Abi-Hassan SP 24 MGSC 2301](https://bit.ly/Abi-Hassan_SP_24_MGSC_2301)





What is Excel?

Excel is a program used to create and edit **spreadsheets**. In Excel, data is organized into rows and columns; this data can be presented and analyzed using Excel's functions, such as pivot tables, charts, formulas, and more.





Why Excel?

Excel is an excellent way to store, organize, and analyze both data and metadata (data about data). Although it is particularly useful for budgeting and finance because many of its functions revolve around numerical data, Excel is used quite often across the disciplines.

In humanities and social science contexts, you might use Excel to pursue research interests, particularly for materials that are provided as spreadsheets (census data, bibliographies, and more).



Common Ways to Use Excel

- Tracking job applications
- Budgeting for events in your personal life
- Collaborative task-tracking (Google Sheets can also be helpful for this)
- Outlining content to be written for a website
- Analyzing data stored in .csv (comma separated value) files
- Collecting and analyzing survey information



Example dataset: Co-op application tracker

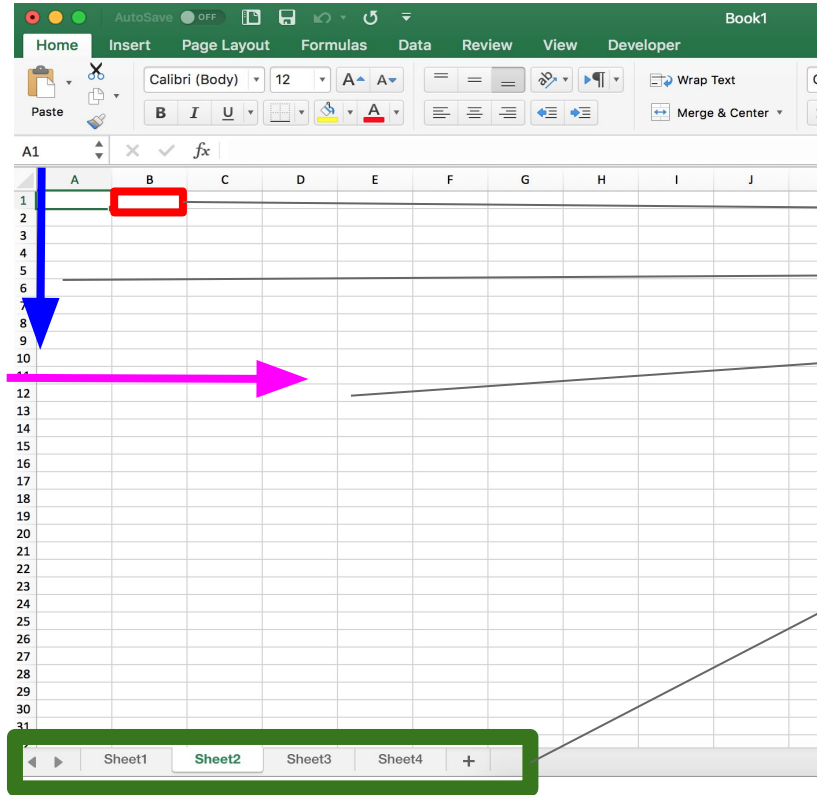
Home Insert Draw Page Layout Formulas Data Review View Acrobat Tell me						
Corbel (Body) 12 A^ A^ B I U % Conditional Formatting Format as Table Cell Styles Cells Editing Analyz Data						
F17						
	A	B	C	D	E	F
1	Position	Organization	Pay (hourly wage)	Application Deadline	Applied (Y/N)	Interview?
2	Research Assistant	Ford's Theatre	18.00	19-Mar	Y	Pending
3	Museum Educator	The Spy Museum	16.00	31-Mar	Y	Declined
4	Digital Intern	White House Historical Association	20.00	22-Feb	Y	Declined
5	Museum Intern	Department of the Interior Museum	15.00	28-Feb	Y	Interview on 10-Mar
6	Library Assistant	DC History Center	19.25	19-Mar	Y	Interview on 15-Mar
7	Historic Preservation Intern	National Park Service	15.00	15-Feb	Y	Interview on 10-Apr
8	Curatorial Assistant	Maryland Historical Society	17.50	1-Apr	N	Pending
9						
10						
11						
12	note: these positions are fictional					

Important Vocabulary

- **Workbook:** the overall Excel file that you are creating
- **Sheet:** the different sheets inside the workbook; these can be renamed
- **Row:** the horizontal and numerical rows
- **Column:** the vertical and alphabetical columns
- **Cell:** the boxes that each have an ID based on their row and column placements (A1, A2, A3, etc).



Anatomy of Excel



CELL



COLUMN



ROW



SHEET



Important Excel Features

- **Functions:** Used to calculate and analyze numerical data, for example with: mean, median, standard deviation, addition, subtraction, and other forms of arithmetic.
- **Tables and Pivot Tables:** Used to filter, analyze, and summarize numerical data, and present different results based on functions and data chosen.
- **Charts:** Used to visualize data with bar charts, scatter plots, and other formats.



How to Select Data

If you have a long dataset, it can be hard to drag your mouse down to the bottom of the dataset. Click

SHIFT + COMMAND/CONTROL + DOWN ARROW
(or whatever direction)

The end of the data will be selected in the direction of the arrow you choose.



Basic Calculations

Using **functions**, you can find the:

- Average (arithmetic mean)
- Mode & Median
- Standard deviation
- Min/max values
- Correlation
- Results for other basic calculations such as addition, subtraction, division, multiplication



Writing Excel Functions

- In an empty cell, type = and then the calculation you want to do:
 - Sum: SUM()
 - Average: AVERAGE()
 - Median: MEDIAN()
 - Standard Deviation: STDEV()
- Select the range to calculate. You can enter the names of the cells or manually select the cells you want included. Then close the brackets.
- You can also write functions referencing other worksheets by using the sheet name and '!'. Example:
 - =AVERAGE(Sheet1!C2:C8)

C	D
Pay (hourly wage)	
\$17.50	
\$20.00	
\$15.00	
\$19.25	
\$16.00	
\$15.00	
\$18.00	=SUM(C2:C8)

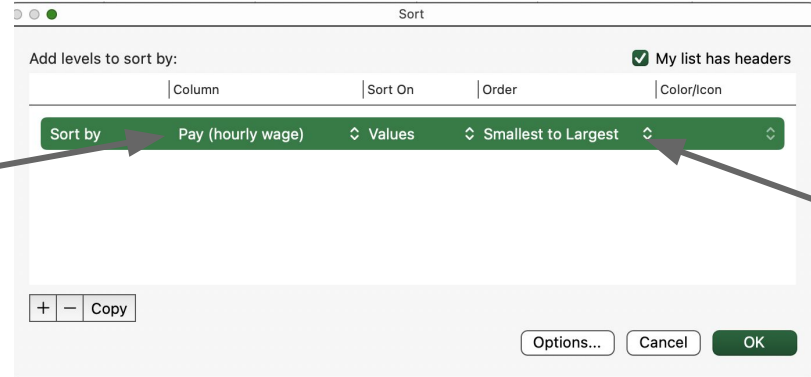
The selected data (C column from rows 2–8)

The function (SUM) with the selected data



Sorting Data

- **Sorting** allows you to organize your data by a certain value
- Select your dataset.
- Select “Sort” under the “Data” tab. Once you click, a pop-up window will appear.
- Choose which column you would like to sort values by, and how you would like to order the sort. The entire dataset will be sorted accordingly.
- If your list has headers (column titles) make sure to tick the right-upper box. Otherwise, Excel will automatically sort the column labels or titles as well.
- You can use the + button to add multiple sort requests.



Select column
from
drop-down

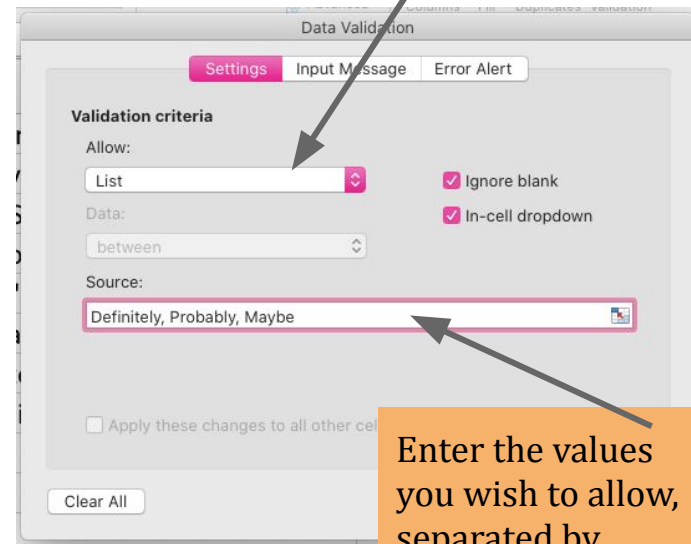
The order will vary
depending on the
variable's category
(number, text)



Adding Data Validation

- **Data validation** allows you to set a limited range of responses (either numbers or words/letters) for a selected group of cells.
- Highlight the cells to which you want to apply the data validation
- Under the “Data” tab Select “Validation”
- Change “Allow” from “Any value” to “List” in the drop-down menu
- Type the responses you want to allow, separated by commas and spaces
- When applying data validation to filled-in cells, Excel will automatically overwrite the cell content. You can avoid this by creating new columns to apply data validation to.

Change from
“Any value” to
“List”



Enter the values
you wish to allow,
separated by
commas and spaces



Adding Conditional Formatting

- Conditional formatting adds automatic color-coding based on your data values
- Highlight the cells to which you want to apply the conditional formatting
- Select “Conditional Formatting” under the “Home” tab and choose from a range of color-coding options
 - Options include Highlighting Rules, Data Bars, Color Scales, and Icon Sets
 - You can visualize numerical variables with data bars (left) or set rules for specific text (right)

Pay (hourly wage)	Interview?
17.50	Pending
20.00	Declined
15.00	Interview on 10-Apr
19.25	Interview on 15-Mar
16.00	Declined
15.00	Interview on 10-Mar
18.00	Pending

Data Bar
visualization

Highlighting
rule



Creating a Table

- **Tables** allow you to present your information in a more polished way. They also create a visual border between your data and the rest of the spreadsheet document.
- Select all the cells that you want included in your table
- Under the “Insert” tab, select “Table”
- You can customize the appearance of your table under the “Table” tab, much as you would in Microsoft Word
- You can still modify your data once it is in a table; although tables make your data look more presentable, they are not a “finished” form

Your Turn: using the same data, insert and customize a table.



Creating a Pivot Table

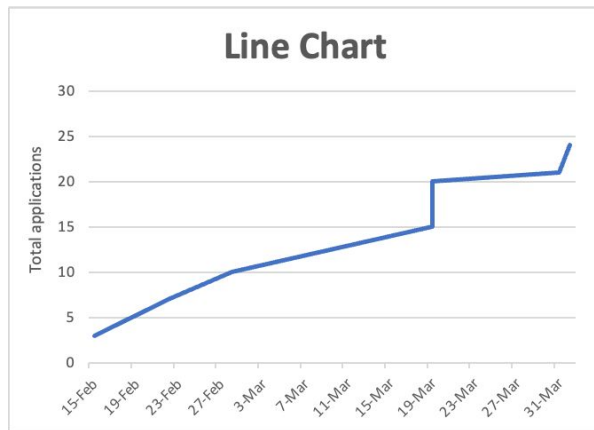
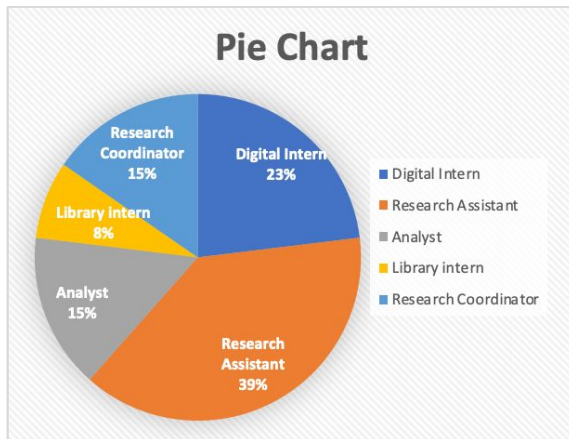
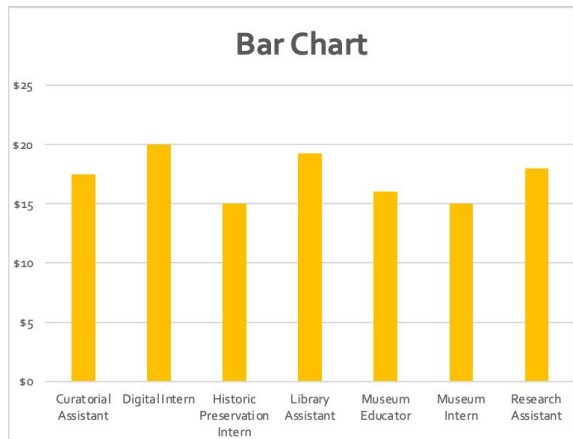
- **Pivot Tables** allow you to present your information in an aggregated polished way and are a powerful means of aggregating and sorting data
- Select all the cells that you want included in your table
- Under the “Insert” tab, select “Pivot Table”
- You can still modify your data once it is in a pivot table; although tables make your data look more presentable, they are not a “finished” form

Your Turn: using the same data, insert and customize a table.



Charts

- While tables represent data or information in rows and columns, a chart is the graphical representation of data in symbols like bars, lines, and slices.
- There are many types of charts you can create with Excel



Creating a Chart

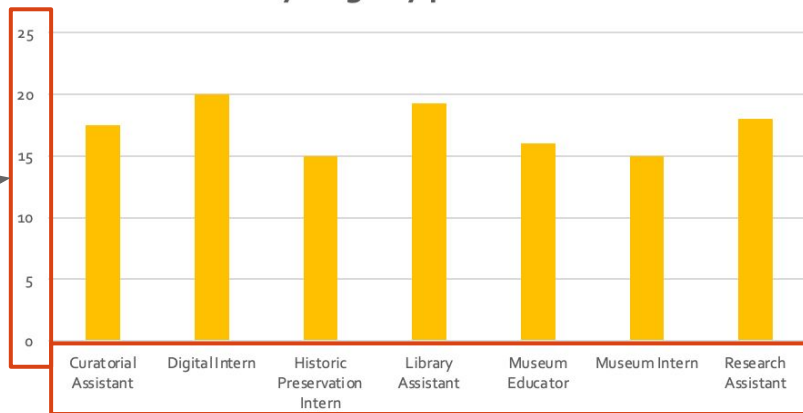
- First, you must **select a chart type** that suits your data:
 - **Bar Chart:** compares parts of a bigger set of data, highlights different categories, or shows change over time. Be careful not to overload your graph, avoid having more than 10 bars.
 - **Pie Chart:** shows relative proportions and percentages of a whole dataset. Best used with small datasets (up to 6 categories).
 - **Line Chart:** for continuous dataset that changes over time. Use it if your dataset is too big for a bar chart, or if you want to visualize trends instead of exact values.
 - **Scatter plots, bubble charts, area charts, etc...** You can learn more [here](#)



Creating a Chart (cont'd)

- Select the columns and variables you would like to include
 - For multiple columns, you may need to move the columns next to each other.
- Go to “Insert” and then “Charts”. Choose the chart type you want.
- If you are creating a bar or line chart, consider what your x-axis and y-axis should be.

Hourly wage by position



y-axis: should be numerical value

x-axis: should contain your categories



Formatting your chart

Once you have your chart, you may want to customize some aspects for more clarity and precision.

- **Format axis:** right-click either axis and select “Format Axis”. Under “Number” you can choose the appropriate category (number, currency, date), the number of decimal places, and the scale for the axis.
- **Add Elements:** under the “Chart Design” toolbar at the top, select “Add Elements” to add or delete chart and axis titles, data labels, gridlines, and legends.
- **Other formatting:** Change colors, font and size under “Chart Design” and “Format”.





Additional Resources

The Internet has a wealth of Excel tutorials. Some particularly useful ones are linked below, including a tutorial for pivot tables, which were not covered in this workshop.

- [Data Validation](#)
- [Conditional Formatting](#)
- [Creating Charts](#)
- [Pivot Tables](#)

Example Dataset 2

A	B	C	D	E
harm_category	retweet_count	reply_count	like_count	quote_count
a	0	0	0	0
a	0	0	1	0
u	116	0	0	0
a	0	0	0	0
a	0	1	1	0
a	0	1	28	0
a	0	0	1	0
u	116	0	0	0
u	116	0	0	0
u	116	0	0	0
u	116	0	0	0
a	36	0	0	0
b	0	1	0	0
e	0	0	1	0



Installing 'Analysis Toolpak'

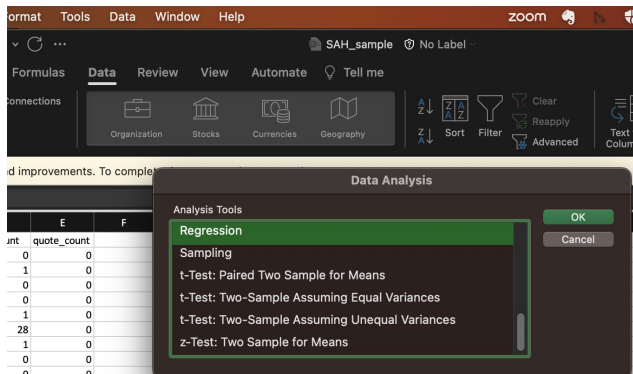
- Analysis Toolpak “provides data analysis tools for statistical and engineering analysis.” It is an Excel add-in that allows for easy statistical analysis like bivariate and multivariate regression. We will also show you how to do regression analysis without the add-in.
- **For MacOS:** click on the “Tools” menu and select “Excel Add-ins”. In the “Add-ins available” box, check the “Analysis ToolPak” box. If you are unable to find this option, search for “Excel Add-ins” under the “Help” menu. If you have an older version of Excel, you may need to go to the Excel options in the “File” menu and find Add-ins there.
- **For Windows:** Click the “File” menu, then select “Options”, then the “Add-ins” category. In the “Manage” box, select “Excel Add-ins” and then click “Go.” The “Add-ins” box will appear, and there you can select “Analysis Toolpak” and click “Ok”.

Compatibility: Should work with most versions of excel and OS software.



Regression Excel

Go to tools -> Data
Analysis



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	harm_category	retweet_count	reply_count	like_count	quote_count								
2	a	0	0	0	0								
3	a	0	0	1	0								
4	u	116	0	0	0								
5	a	0	0	0	0								
6	a	0	1	1	0								
7	a	0	1	28	0	E5							
8	a	0	0	1	0								
9	u	116	0	0	0								
10	u	116	0	0	0								
11	u	116	0	0	0								
12	u	116	0	0	0								
13	a	36	0	0	0								
14	b	0	1	0	0								
15	e	0	0	1	0								
16	a	13	0	0	0								
17	a	1	0	0	0								
18	a	1	0	4	0								
19	u	0	0	0	0								
20	b	1	0	0	0								
21	b	1	0	2	0								
22	b	0	1	5	0								
23	u	0	0	1	0								
24	u	3	0	0	0								
25	a	1	0	0	0								
26	a	1	1	19	0								
27	e	4	0	0	0								
28	e	4	0	0	0								
29	e	4	0	0	0								
30	e	3	0	0	0								
31	e	3	0	0	0								
32	e	3	0	5	0								
33	b	122	0	0	0								

Regression

Input

Input Y Range:

Input X Range:

☐ Labels ☐ Constant is Zero

☐ Confidence Level: %

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel



Excel Regression Output

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.6347765							
R Square	0.40294121							
Adjusted R S	0.37997741							
Standard Error	0.30710846							
Observations	28							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	1.6549371	1.6549371	17.5468002	0.00028524			
Residual	26	2.45220576	0.0931561					
Total	27	4.10714286						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	0.0893832	0.06182028	1.44585561	0.16016513	-0.0376902	0.21645661	-0.0376902	0.21645661
X Variable 1	0.03963921	0.00946294	4.1888901	0.00028524	0.02018786	0.05909056	0.02018786	0.05909056



Excel Horror Stories

How an Excel Error Cost JP Morgan \$6 Billion

In 2012, JP Morgan Chase, a leading financial services firm, faced a \$6 billion loss, due to an Excel error in their Value-at-Risk (VaR) model.

4. Fidelity Investment's Minus Symbol Blunder

In 1994, Fidelity Investments estimated they would make a \$4.32 per share dividend distribution on their Magellan fund by the end of the year and promised to pay their shareholders accordingly. In January of 1995 though, the financial institution was forced to cancel the said dividend distribution after discovering that their estimation was incorrect. Massively incorrect.

How did this happen? An employee simply forgot to put in a minus sign.

While transferring financial records onto an Excel spreadsheet, a tax accountant neglected to put a minus sign on the fund's net capital loss of \$1.3 billion. The loss was then calculated as a net capital gain, which resulted in dividend estimates being off by a staggering \$2.6 billion.



R - A Statistical Computing Language

- What is R?
- Advantages of R over Excel
- Disadvantages of R over Excel
- Demo of R and Excel!



Thank you!

If you have any questions, contact us at nulab.info@gmail.com

Schedule an appointment with DITI:

<https://docs.google.com/forms/d/e/1FAIpQLSd2TcUF8IhgUat7j2J3BXG54zNIpz6EBdiaIoqPqOwBjkbPIA/viewform>

We'd love your feedback! Please fill out a short survey here:

<https://bit.ly/diti-feedback>

Slides, data, and code available at the link below:

 https://bit.ly/Abi-Hassan_SP_24_MGSC_2301
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!