JOHNS HOPKINS UNIVERSITY
Pre-College Program
AS.110.100 Data Analytics Workshop

Excel Notes

Aaron Zoll

How to use these notes. These notes follow the Excel videos, one section per video, and are optional in the same way the lecture notes are. The first page is a quick reference with every function and shortcut we use, which is worth keeping open while you work on the guided practice and your project. Excel is best learned by doing, so I invite you to open a blank workbook and follow along.

TL;DR (Quick reference) Every formula starts with `=`. A range like `A4:A7` means the cells `A4` through `A7`, and a `$` in a reference (`$D$4`) keeps it from shifting when the formula is copied.

Function	What it returns	Math
`=SUM(A4:A7)`	Sum of the cells	$\sum x_k$
`=COUNT(A4:A7)`	How many cells hold numbers	N
`=AVERAGE(A4:A7)`	Mean (skips blanks)	$\bar{x}$
`=MEDIAN(A4:A7)`	Median	$\hat{x}$
`=MODE.SNGL(A4:A7)`	Most common value	$\tilde{x}$
`=MAX(...)`, `=MIN(...)`	Largest, smallest value	
`=VAR.P(A4:A7)`	Variance (divides by N)	σ^2
`=STDEV.P(A4:A7)`	Standard deviation	σ
`=SQRT(x)`	Square root	$\sqrt{x}$
`=STDEV.P(r)/SQRT(COUNT(r))`	Standard error of the mean of a range `r`	$\frac{\sigma}{\sqrt{N}}$
`=CORREL(A2:A50, B2:B50)`	Correlation of two paired columns	$\rho_{x,y}$
`=RSQ(B2:B50, A2:A50)`	R^2 of the line of best fit	$\rho^2_{x,y}$
`=IF(test, a, b)`	`a` if the test is true, `b` if not	
`=ABS(x)`	Absolute value	$\lvert x \rvert$
`=TRIM(C2)`	Text with extra spaces removed	
`=UPPER(C2)`, `=LOWER(C2)`	Text in all capitals, all lowercase	

Keys	Action
Ctrl + arrow	Jump to the last filled cell in that direction
Ctrl + **Shift** + arrow	Highlight to the last filled cell in that direction
Ctrl + **D**	Fill the top cell's formula down the highlighted cells
Ctrl + **C**, **Ctrl** + **V**, **Ctrl** + **Z**	Copy, paste, undo
Double click a cell's bottom right corner	Fill a formula down as far as the data next to it goes

- **Charts**: highlight the data, then **Insert** → **Recommended Charts** → **All Charts**. Clustered Column for bar charts, Histogram for one numerical column, Scatter for two.
- Add titles, axis titles (with units), error bars, and trendlines with the + button next to a chart (Mac: **Chart Design** → **Add Chart Element**).
- **Error bars**: compute $2 \cdot \text{SEM}$ for each group in cells, then **Error Bars** → **More Options** → **Custom** → **Specify Value**, and point both boxes at those cells. Never use Excel's built in "Standard Error" option.

- On a Mac, use **Cmd** wherever these notes say **Ctrl**. Stuck? See the troubleshooting appendix.

Contents

1 Getting Started (Day 1) 3
1.1 Excel on a Mac . . . 3
1.2 Cells and Data Entry . . . 3
1.3 On a Cell versus In a Cell . . . 4
1.4 Column Width and Wrapping Text . . . 4

2 Navigating and Entering Data (Day 1) 4
2.1 Keyboard Navigation . . . 4
2.2 Creating a Series . . . 4

3 Functions (Day 2) 5
3.1 Copying Formulas . . . 5

4 Cleaning Data (Day 3) 6
4.1 Removing Blank Rows . . . 6
4.2 Precision . . . 6
4.3 Extra Spaces and Capitalization . . . 6
4.4 Typos and Missing Values . . . 6
4.5 Dates . . . 6
4.6 Sorting . . . 6
4.7 Choosing Variables . . . 7

5 Mean and Variance (Day 4) 7
5.1 Small Datasets . . . 7
5.2 Large Datasets . . . 7

6 Bar Charts and Error Bars (Day 5) 8
6.1 Bar Charts . . . 8
6.2 Error Bars . . . 8
6.3 More Than One Series . . . 10

7 Histograms (Day 6) 11

8 Scatter Plots and Trendlines (Day 6) 12

A Troubleshooting 13
A.1 Error Values in Cells . . . 13
A.2 Formulas . . . 14
A.3 Cleaning and Sorting . . . 14
A.4 Charts . . . 14
A.5 Error Bars . . . 15

1 Getting Started (Day 1)

For this course we start each project from a **Blank workbook**. Almost everything in these notes also works in Google Sheets, with some menus in different places. The one big exception is error bars: Google Sheets can only give every bar the same kind of error bar (a constant, a percent, or a standard deviation), not a value we computed for each group, so make your error bar charts in Excel.

1.1 Excel on a Mac

The videos use Excel on Windows. Excel on a Mac has the same functions and nearly the same menus, and the formulas are identical. The main differences are the keys and a few chart menus.

Action	Windows	Mac
Copy, paste, undo	**Ctrl** + **C**, **V**, **Z**	**Cmd** + **C**, **V**, **Z**
Jump to the last filled cell	**Ctrl** + arrow	**Cmd** + arrow
Highlight to the last filled cell	**Ctrl** + **Shift** + arrow	**Cmd** + **Shift** + arrow
Fill down	**Ctrl** + **D**	**Cmd** + **D** (or **Ctrl** + **D**)
Select cells that are not next to each other	**Ctrl** + click	**Cmd** + click
Right click	right mouse button	**Ctrl** + click, or two finger click
Format the selected cell or chart part	**Ctrl** + **1**	**Cmd** + **1**
Chart titles, error bars, trendlines	+ button next to the chart	**Chart Design** → **Add Chart Element**

Throughout these notes, whenever you see **Ctrl**, a Mac user should press **Cmd**. If a shortcut does nothing on a Mac, check whether macOS is using it for something else (for example, **Ctrl** + arrow switches desktops), and use the Mac version in the table.

1.2 Cells and Data Entry

A sheet is a grid of cells. Columns are labeled by letters, going $A, \ldots, Z, AA, \ldots, ZZ, AAA$ and on to XFD, and rows are labeled by numbers, going past one million. The cell in column D and row 6 is called `D6`, and the name of the selected cell is shown in the top left corner. We will never need anywhere near this much room.

A cell can hold three types of data.

1. **Text**, such as a column title.
2. **Numbers**, with as many decimal places as we like. If a number is wider than its column, Excel displays it rounded (but still remembers every digit).
3. **Formulas**, which always begin with `=`. Typing `=1+1` returns 2 and `=4*9` returns 36. A formula can also use other cells: `=B8+B9` returns the sum of the values in `B8` and `B9`.

Formulas are what make Excel excel. By default, text is left aligned and numbers are right aligned, which is a quick way to spot a number Excel has accidentally stored as text. The **Home** tab has the same formatting options as Word: font, size, bold, color, and horizontal and vertical alignment.

1.3 On a Cell versus In a Cell

The formula bar above the sheet shows what is actually inside the selected cell. For a cell containing `=4*9`, the cell displays 36 but the formula bar shows `=4*9`.

- Clicking a cell once puts us **on** the cell. We can change its formatting, and typing replaces its contents while keeping the formatting.
- Double clicking a cell, or clicking in the formula bar, puts us **in** the cell. Now we can edit the contents, for example adding a digit to a number.

1.4 Column Width and Wrapping Text

Long text spills into neighboring columns, and long numbers get rounded. We can drag the boundary between two column letters to change the width, or double click the boundary to fit the column to its contents. Rows work the same way. To keep long text inside one cell, select it and click **Home** $\rightarrow$ **Wrap Text**.

2 Navigating and Entering Data (Day 1)

When entering data, always include units in the column titles, such as "Height (in)" and "Weight (lb)." Future you will thank you. The **Borders** menu on the **Home** tab adds lines, such as a thick bottom border under the titles, to make a dataset easier to read.

2.1 Keyboard Navigation

Keys	Action
Arrow keys	Move one cell
Enter / **Shift** + **Enter**	Move down / up
Tab / **Shift** + **Tab**	Move right / left
Ctrl + arrow	Jump to the last filled cell in that direction
Ctrl + **Shift** + arrow	Highlight to the last filled cell in that direction
Ctrl + **Z**	Undo

Ctrl + arrow is especially useful for large datasets. From the top left corner of a dataset, **Ctrl** + **Shift** + **Down** followed by **Ctrl** + **Shift** + **Right** highlights the whole dataset, so we can format it all at once. Clicking a column letter or row number highlights the entire column or row.

2.2 Creating a Series

To enter $1, 2, \ldots, 6$ we can just type them. To enter $1, 2, \ldots, 1000$ we let Excel do it. If we type 1 and 2 in two cells, highlight both, and drag the bottom right corner of the selection down, Excel recognizes the pattern and continues it. This works for other step sizes too.

For long lists, it is easier to type the first value, select its cell, and choose **Home** $\rightarrow$ **Fill** $\rightarrow$ **Series**. Choose **Columns**, a **Linear** type, a **Step value**, and a **Stop value**. For example, a first value of 3 with step value 7 and stop value 1000 gives $3, 10, 17, \ldots, 997$, stopping at the last value that does not exceed 1000.

3 Functions (Day 2)

Every formula begins with `=` and follows the order of operations, including exponents such as `=2^3`. Beyond arithmetic, Excel has many built in functions. Typing `=` followed by a letter shows a list of them, and typing the open parenthesis shows what the function expects as input. The quick reference on the first page lists every function we use in this course.

Function	Returns
`=SUM(A4:A7)`	The sum of the cells `A4` through `A7`
`=AVERAGE(A4:A7)`	The mean of those cells, skipping blanks
`=COUNT(A4:A7)`	The number of numerical cells
`=ABS(-4)`	4
`=IF(7>3, "True!", "False!")`	`True!`

`IF` takes a logical test, a value to return if the test is true, and a value to return if it is false. Text values must be in quotation marks. The test `7>13` would return `False!`.

3.1 Copying Formulas

Here is where Excel really saves time. When we copy a formula to another cell, Excel shifts the cell references with it. If `D4` contains `=C4+18` and we drag the bottom right corner of `D4` down, then `D5` contains `=C5+18`, `D6` contains `=C6+18`, and so on. Dragging to the right shifts the column instead, so `=D4*7-3` becomes `=E4*7-3`. If we do *not* want a reference to shift, we lock it with dollar signs, as in `$D$4` (we will need this for the variance).

Dragging is tedious for 1000 rows. A faster method:

1. In the row just below the last data point of the column we are filling, type a **stopper**, such as the word `stop`. This keeps **Ctrl** + **Down** from jumping to the bottom of the sheet.
2. Select the cell with the formula and press **Ctrl** + **C**.
3. Press **Ctrl** + **Shift** + **Down** to highlight down to the stopper, then release **Ctrl** and press **Shift** + **Up** to leave the stopper out.
4. Press **Ctrl** + **D** to fill the formula down.

Double clicking the bottom right corner of a formula cell also fills the formula down as far as the neighboring data goes.

Example 3.1

Suppose column `M` contains `=(L4*3-9)/4` filled down beside the list $1, 2, \ldots, 1000$ *in column `L`. In column `N` we enter*

```
=IF(M4>400, "Bigger than 400", "Still too small")
```

and fill it down. Scrolling to where the text changes shows where the values in column `M` pass 400. *(As a sanity check, solve* $\frac{3L-9}{4} > 400$ *by hand, like on Day 1 of the lecture notes. Where should it switch?)*

4 Cleaning Data (Day 3)

Raw data is cluttered, and cleaning it takes a good amount of time. It pays off when we start the analysis. A good first step is to select all the columns and double click a column boundary so every column fits its contents. The lecture notes for Day 3 talk about *what* to remove (outliers); this section is about *how*.

4.1 Removing Blank Rows

Highlighting blank rows by hand is unreliable for anything but small datasets. Instead, click **Home** → **Find & Select** → **Go To Special** → **Blanks**, then **Home** → **Delete** → **Delete Sheet Rows**.

Be careful: **Blanks** also selects single blank cells, and deleting their rows deletes the rest of the data in those rows. If a dataset has scattered blank cells, it can be better to sort the column so the blanks collect at the top or bottom, or to leave the blanks until the analysis. When comparing two columns, we only need to drop rows where one of those two columns is blank.

4.2 Precision

Values such as stock prices may come with far more decimal places than we need. Select the columns and use **Increase Decimal** or **Decrease Decimal** in the **Number** group of the **Home** tab.

4.3 Extra Spaces and Capitalization

To clean a text column, right click the column letter to its right and choose **Insert** to make a helper column. In the helper column enter `=TRIM(C2)`, which removes the extra spaces from the text in `C2`, and fill it down. Similarly, `=UPPER(C2)` or `=LOWER(C2)` makes the capitalization consistent.

The helper column depends on the original column, so deleting the original gives an error. Instead, copy the helper column, right click the original column, choose the paste option **Values**, and then delete the helper column.

This matters for analysis. If one entry reads "Amazon.com, Inc." and another has extra spaces, Excel treats them as different names, and an average by company would quietly split Amazon into two companies.

4.4 Typos and Missing Values

Home → **Find & Select** → **Replace** replaces every occurrence of one string with another, for example every "Twiiter" with "Twitter." If some data is selected, Find only searches the selection. The same tool can replace placeholder text for missing values, such as `nan`, with a blank.

4.5 Dates

Some dates may appear as five digit numbers. Excel stores a date as the number of days since January 1, 1900, so 44379 is July 2, 2021. To make all dates consistent, select them and choose **Short Date** from the **Number** format menu.

4.6 Sorting

To sort, highlight all the data including the titles and choose **Home** → **Sort & Filter** → **Custom Sort**. Pick the column to sort by and the order, such as A to Z. Excel sorts by that one column but

moves each entire row with it, so every data point stays together. (If you only highlight one column, Excel will warn you; sorting just that column would scramble whose data is whose.)

4.7 Choosing Variables

A dataset may have far more variables than we need. Part of cleaning is deciding which variables to keep. A dataset with 126 samples and around 100 variables might be reduced to GPA, gender, and a few breakfast related columns. Deleting every row with a blank *anywhere* in this dataset dropped it to about 36 samples, so only remove rows with blanks in the variables you actually use.

5 Mean and Variance (Day 4)

5.1 Small Datasets

For data in cells `A4` through `A7`, the total is `=A4+A5+A6+A7`, or more simply `=SUM(A4:A7)`. To select the range quickly, type `=SUM(`, click `A4`, and press **Ctrl** + **Shift** + **Down**. The mean is any of `=SUM(A4:A7)/4`, `=SUM(A4:A7)/COUNT(A4:A7)`, or `=AVERAGE(A4:A7)`. Recall the mean is a weighted average with the same weight, here $\frac{1}{4}$, on each data point.

The mean does not describe the spread. To compute the variance by hand, building the same table as in the Day 4 lecture notes, with the mean in `D4`:

1. In `E4` enter `=A4-$D$4` and fill down. The dollar signs lock the reference to `D4`, so it does not shift when the formula is copied. This column is $x_k - \bar{x}$.
2. In `F4` enter `=E4^2` and fill down. Squaring makes each deviation positive. This column is $(x_k - \bar{x})^2$.
3. Sum column `F` to get the total squared error, $\sum(x_k - \bar{x})^2$.
4. Divide the total squared error by the count to get σ^2.

This matches Excel's built in `=VAR.P(A4:A7)`, which for this randomly generated data gives a mean of about 0.70 and a variance of about 0.098. It is worth doing it by hand once, so that `VAR.P` is not a black box.

Remark 1

Excel also has `VAR.S`, which divides the total squared error by one less than the count. This is used when the data is a sample from a larger population, such as measuring a few hundred newborns instead of every newborn in a state. The number $N-1$ is called the degrees of freedom. In this course we always use `VAR.P` (and `STDEV.P`).

5.2 Large Datasets

The same steps work for any size dataset. For long columns, use a stopper and **Ctrl** + **D** to fill the formulas down, as in Section 3.1.

The data here are random numbers between 0 and 1, each equally likely, so we expect the mean to approach 0.5. With 153 data points the mean is within about 0.015 of 0.5, and with over 3000 it is within about 0.005. The variance, however, stays between about 0.08 and 0.1 no matter how

much data we add. More data brings the mean closer to its expected value, but it does not change how spread out the individual data points are.

To bring the variance down, we average. Split the data into columns, take the mean of each column, and then take the mean and the variance of those means. With 10 values per column, the variance of the means is about 0.01, roughly 10 times smaller. With 563 values per column, the mean of the means is very close to 0.5 and the variance of the means is far smaller still. Each column mean combines values above and below 0.5, so the means fall closer to 0.5 than individual values do. This is exactly the standard error of the mean from Day 4 (the variance of an average of N values is $\frac{\sigma^2}{N}$), and the central limit theorem from Day 7: a histogram of those column means would look like a bell curve.

6 Bar Charts and Error Bars (Day 5)

6.1 Bar Charts

A bar chart is for *categorical* data: one bar per group, and the height of a bar is a number describing that group, usually its mean. To chart data, highlight it and use the **Quick Analysis** button that appears at the bottom right of the selection, or go to **Insert** → **Recommended Charts**. The recommendations are not always the best choice, so it helps to look under **All Charts**. A **Clustered Column** chart is the bar chart we use in this course. 3D charts rarely add anything, so keep it simple.

Some things to check:

- **Scale.** Charting the Open, High, Low, Close, and Volume columns together makes the first four invisible, since Volume is far larger. Leave out columns on a very different scale.
- **Labels.** To label the horizontal axis by category, include the category column, here Quarter, in the selection. To select columns that are not next to each other, hold **Ctrl**.
- **Titles.** Use the + button next to the chart to add a chart title and axis titles with units. The legend can be moved or renamed there too.
- **Grouping.** Charting Amazon, Google, and Twitter in one chart lumps three companies together under repeated quarters. Make one chart per company instead.

On a Mac there is no + button next to the chart. Instead, click the chart and use **Chart Design** → **Add Chart Element**, which has the same list: Axis Titles, Chart Title, Error Bars, Legend, Trendline, and so on. On either system, double clicking any part of a chart (a bar, an axis, the title) opens a **Format** pane on the right side of the screen with every option for that part.

6.2 Error Bars

The video compares the average closing price of three companies. To see every step with numbers you can check, let's instead walk through a small example you can type in yourself: the plant heights from Day 7 of the lecture notes. Each error bar will reach $2 \cdot \text{SEM}$ above and below its bar (see Day 7 of the lecture notes for why 2).

The most important thing to know before starting: **Excel will not compute the error bars for you**. Its built in options (**Standard Error**, **Standard Deviation**, **Percentage**, **Fixed Value**) do not use the data in each group. For example, Excel's "Standard Error" option computes the standard error of the *bar heights themselves*, which is meaningless here, and gives every bar the same size error bar. We compute the error bars in cells ourselves and point Excel at those cells.

Step 1: Enter the data. One column per group, with the units in the title.

Step 2: Build a summary table. Next to the data, make one row per group and one column per statistic, as below. Type the formulas in row 2, then fill them down to row 3 and change `A` to `B` in row 3 (or type them again).

	A	B	C	D	E	F	G	H
1	**No fert. (cm)**	**Fert. (cm)**		**Group**	**Mean**	**SD**	**SEM**	**2 SEM**
2	7.2	8.3		No fertilizer	7.50	0.432	0.176	0.353
3	6.8	7.9		Fertilizer	8.33	0.315	0.128	0.257
4	7.9	8.8						
5	7.4	8.0						
6	8.1	8.6						
7	7.6	8.4						

Cell	Formula	Meaning
`E2`	`=AVERAGE(A2:A7)`	$\bar{x}$
`F2`	`=STDEV.P(A2:A7)`	σ
`G2`	`=F2/SQRT(COUNT(A2:A7))`	SEM $= \frac{\sigma}{\sqrt{N}}$
`H2`	`=2*G2`	how far each error bar reaches

Using `COUNT` instead of typing 6 means the formula keeps working if you add or remove data later.

Step 3: Make the chart. Highlight `D1:E3`, the group names and the means (and nothing else). Then **Insert** → **Recommended Charts** → **All Charts** → **Clustered Column**. On a Mac, **Insert** → the column chart icon → **Clustered Column**. Because the group names were included, they label the bars automatically.

Step 4: Add the error bars (Windows).

1. Click the chart, then + → hover over **Error Bars** → click the small arrow → **More Options**. The **Format Error Bars** pane opens on the right.
2. Under **Error Bar Options**, set **Direction** to **Both** and **End Style** to **Cap**.
3. Under **Error Amount**, choose **Custom** and click **Specify Value**.
4. A small box appears with a **Positive Error Value** and a **Negative Error Value**, each containing `={1}`. Click in the Positive box, *delete everything in it*, and then highlight `H2:H3` on the sheet.
5. Do the same for the Negative box, also with `H2:H3`, and click **OK**.

Step 4: Add the error bars (Mac).

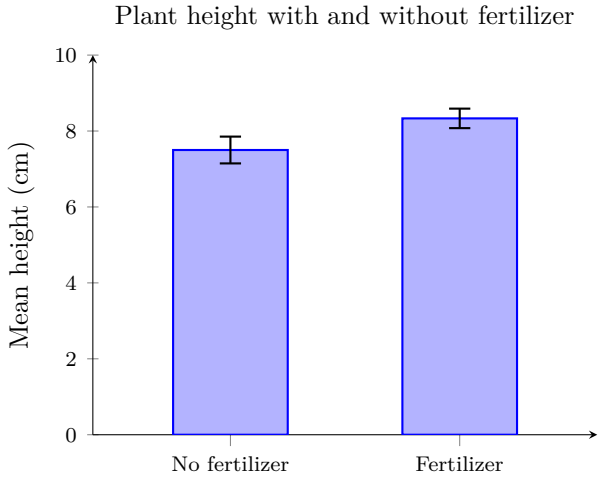
1. Click the chart, then **Chart Design** → **Add Chart Element** → **Error Bars** → **More Error Bars Options**. The **Format Error Bars** pane opens on the right.

2. If the pane does not show the error bar options, click the bar chart icon at the top of the pane (**Error Bar Options**). Set **Direction** to **Both** and **End Style** to **Cap**.

3. Under **Error Amount**, choose **Custom** and click **Specify Value**.

4. In the box that appears, clear `={1}` from the **Positive Error Value**, then highlight `H2:H3` on the sheet. Do the same for the **Negative Error Value**, and click **OK**.

Step 5: Finish the chart. Add a chart title ("Plant height with and without fertilizer") and a vertical axis title with units ("Mean height (cm)"), and delete the legend, since there is only one series. Then check the result against the table. The left error bar should run from about $7.5 - 0.35 = 7.15$ to 7.85, and the right one from about 8.08 to 8.59. If the error bars are all the same size, something went wrong; see the troubleshooting appendix.



The finished chart. The error bars do not overlap, so this (small, made up) experiment gives evidence that the fertilizer helps.

6.3 More Than One Series

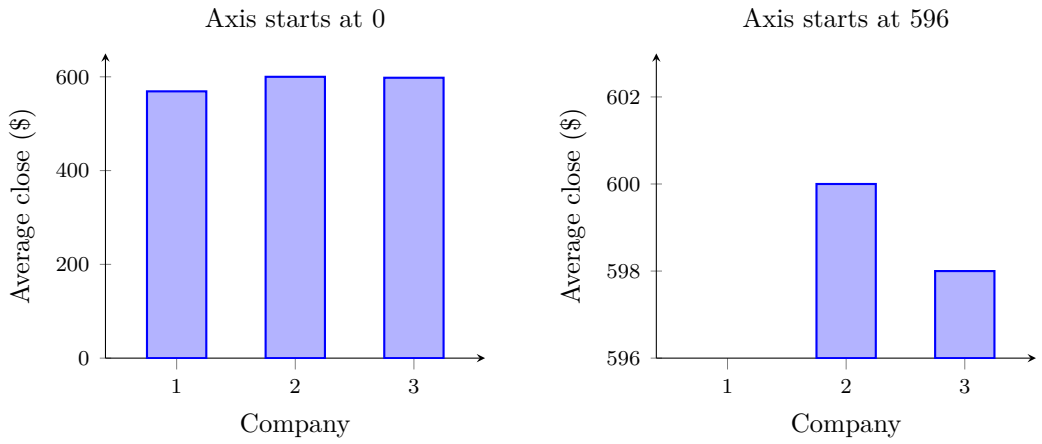
Sometimes each category has two bars, such as average salary by department, split into two age groups. Highlight the category column and both columns of means, and Excel draws a clustered chart with two colors (two **series**) and a legend. Each series gets its own error bars, and each one needs its own **Specify Value**:

- Add error bars as above. Excel asks which series to add them to (or adds them to all of them).
- To switch to the other series, click one of its bars, or pick it from the dropdown at the top of the Format pane (**Error Bar Options** → the small arrow; on a Mac, the **Format** tab's **Current Selection** dropdown also works).
- Each series' Specify Value should point at *that series'* column of $2 \cdot \text{SEM}$ values.

Lay out the summary table with the same rows in the same order for every series, so the error bars line up with the right bars.

Remark 1

The vertical axis should start at 0. With means of 569, 600, and 598, setting the axis to run from 596 to 603 makes the second bar look twice as tall as the third, even though the two are nearly identical. Narrowing the axis like this misrepresents the data, and it is one of the most common ways charts mislead people (keep an eye out for it in the news).



The same averages. On the right, Company 2 looks about twice as large as Company 3, and Company 1 disappears entirely.

7 Histograms (Day 6)

A histogram shows how numerical data is distributed by counting how many data points fall in each range, or **bin**. Select a column (just the one column of numbers, with or without its title) and choose **Histogram** from the recommended charts, or **Insert** → **Insert Statistic Chart** → **Histogram**. On a Mac, it is **Insert** → the statistical chart icon (it looks like a small histogram) → **Histogram**.

The default bins may be too wide, for example only four bins with awkward edges like 63.06 to 353.06. Double click the horizontal axis (or right click it and choose **Format Axis**) and under **Axis Options** → **Bins** set a **Bin width**, such as 50. There are also options for the number of bins and for overflow and underflow bins. The steps are the same on a Mac.

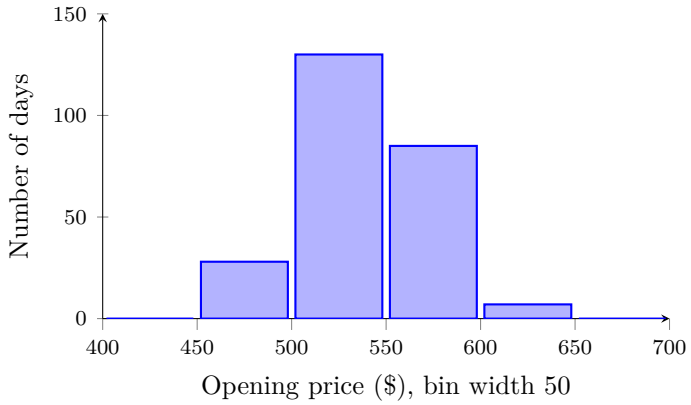
The bins start at the smallest data point, so two histograms of different columns will usually have different bin edges. A workaround to make the bins line up:

1. Add two fake data points to each column, such as 0 and 1050, and color them red so we remember they are not real data.

2. Set the bin width to 50. The bins are now 0 to 50, 50 to 100, and so on, for every column.

3. Select the first and last bars, which contain only the fake points, and set them to **No fill**.

The fake points still add one to the count of the first and last bins, so be careful when reading those bins. Reduce the decimal places on the axis labels under **Number**.



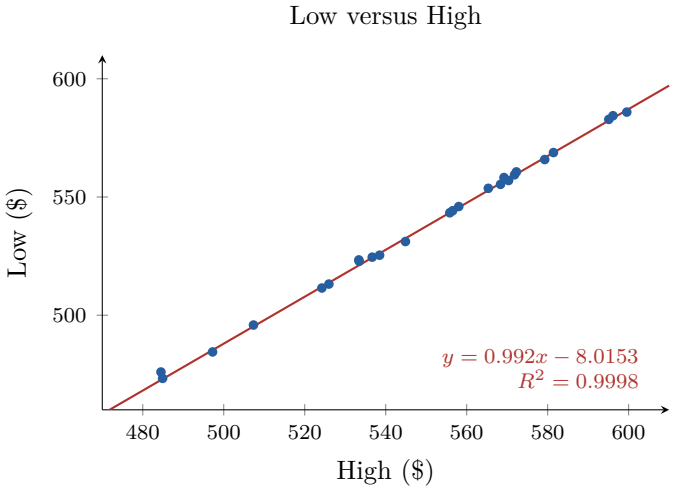
A histogram of (made up) opening prices over 250 days

The histograms of the Open and High columns look very similar. This does *not* mean the two columns are related. A histogram has no order: shuffling the data in a column gives the same histogram. To look for a relationship between two columns we need a scatter plot.

8 Scatter Plots and Trendlines (Day 6)

In the stock data, the low and high values come in pairs: a low of 477 goes with a high of 482, and low values in the 500s go with high values in the 500s. This suggests that as one increases, so does the other. A scatter plot shows this.

Highlight the two columns and insert a **Scatter** chart. Check that Excel plots one column against the other, and not both columns against the data point number. Excel puts the leftmost selected column on the horizontal axis. Add axis titles, and title the chart in the form "y variable versus x variable."



A scatter plot with a linear trendline (points drawn to look like the stock example)

Use a **Scatter** chart, not a **Line** chart. A line chart treats the horizontal values as category names, spaced evenly in the order they appear, so the picture (and any trendline) is wrong whenever the x values are not already evenly spaced and sorted.

The points lie close to a straight line, which indicates a strong correlation. To fit a line of best fit, click + → **Trendline** → **More Options**. On a Mac, use **Chart Design** → **Add Chart Element** → **Trendline** → **More Trendline Options**. Under the trendline options in the Format pane:

- Choose **Linear**. Other types are available.
- Use **Forecast** to extend the line forward or backward. Be careful when extrapolating, since data does not always continue the same way.
- Check **Display Equation on chart** and **Display R-squared value on chart**.

For this data the trendline is $y = 0.992x - 8.0153$ with $R^2 = 0.9998$. A slope near 1 means that when one value increases by 1, the other increases by almost exactly 1. An R^2 this close to 1 means the two columns are almost perfectly correlated, nearly as if we had plotted a column against itself. Recall that any data has a correlation of 1 with itself.

You do not need a chart to get these numbers: `=CORREL(A2:A50, B2:B50)` gives the correlation ρ of two columns directly, and `=RSQ(B2:B50, A2:A50)` gives $R^2 = \rho^2$. It is still worth making the plot, since correlation only measures how *line-like* the data is (see the remark on Day 5 of the lecture notes about $y = x^2$).

Not every point falls on the line. The vertical distance from a point to the line is its **residual**, and the trendline is the line that makes the sum of the squared residuals as small as possible. In a guided practice you will compute R^2 two ways: from the covariance and variances without graphing, and from the trendline and its residuals.

A Troubleshooting

When something goes wrong in Excel, it is almost always one of a handful of things. Find the symptom below. If none of these fix it, undo (**Ctrl** + **Z**) back to the last point where things worked and try again, one step at a time.

A.1 Error Values in Cells

Shows	What it means	Usual fix
`#####`	The column is too narrow to show the number.	Double click the column boundary to widen it. Nothing is wrong with the value.
`#DIV/0!`	Dividing by zero, often an `AVERAGE` or SEM of a range with no numbers in it.	Check the range actually covers the data, and that the data are numbers, not text.
`#VALUE!`	Arithmetic on text, such as `=A2*2` where `A2` holds a word or a number stored as text.	Fix the cell (see "numbers stored as text" below).
`#NAME?`	Excel does not recognize a word in the formula: a misspelled function (`=AVERGE`), or text without quotation marks in an `IF`.	Check the spelling, and put text in quotation marks.
`#REF!`	The formula points at a cell that no longer exists, usually because a column or row it used was deleted.	Undo the deletion. Before deleting an original column (as with `TRIM`), paste the helper column as **Values**.
`#N/A`	A value is not available, for example a lookup that found nothing.	Check the value you are looking for actually appears.

A.2 Formulas

- **The cell shows the formula itself instead of an answer.** Either the formula is missing its `=` (or has a space before it), or the cell is formatted as **Text**. Change the format to **General** in the **Number** menu, then click in the cell and press **Enter**.
- **Numbers stored as text.** The numbers are left aligned, there is a small green triangle in the corner of the cell, and `SUM` or `AVERAGE` ignores them. This often happens with downloaded data. Select the cells, click the yellow warning symbol that appears, and choose **Convert to Number**.
- **A filled down formula gives nonsense after the first row.** A reference shifted that should not have. In the variance example, `=A4-D4` filled down becomes `=A5-D5`, and `D5` is empty. Lock it with dollar signs, `=A4-$D$4`, and fill down again. Click a few of the filled cells and read the formula bar to check.
- **My variance (or standard deviation) does not match a friend's.** One of you used `VAR.S` or `STDEV.S` (or the older `VAR` and `STDEV`, which are the same as the `.S` versions). In this course use `VAR.P` and `STDEV.P`.
- **The average seems off.** `AVERAGE` skips blank cells and text, but it *counts* zeros. If missing values were entered as 0, they drag the average down. Replace them with blanks.
- **Ctrl + Shift + Down stops early, or jumps to the very bottom of the sheet.** It stops at the first blank cell, and if it starts at the last filled cell it jumps to row 1,048,576. Remove blank rows first, or use a stopper as in Section 3.1.
- **Dates turned into five digit numbers (or numbers turned into dates).** Change the format in the **Number** menu to **Short Date** (or back to **General**).

A.3 Cleaning and Sorting

- **Sorting scrambled the data.** Only one column was sorted, so its values are no longer next to the rest of their row. Undo immediately, highlight *all* the columns (including titles), and sort again with **Custom Sort**. If **My data has headers** is not checked, the titles get sorted into the data.
- **Deleting blank rows deleted real data. Go To Special → Blanks** selects every blank cell, and deleting its row removes the rest of that row. Undo, and only delete rows with blanks in the columns you actually use.
- **I cannot find Go To Special on a Mac.** Try **Home → Find & Select → Go To Special**. On older versions, use **Edit → Find → Go To**, then click **Special**.
- **Find and Replace changed too much.** Replace works inside words too, so replacing every "Inc" with "Inc." also changes "Inc." into "Inc..". Check **Match entire cell contents** under **Options**, or select just the column first.

A.4 Charts

- **The horizontal axis says $1, 2, 3$ instead of the group names.** The group names were not included in the selection. Right click the chart, choose **Select Data**, and under **Horizontal (Category) Axis Labels** click **Edit** and highlight the cells with the names.

- **The group names show up as their own bars.** The names were numbers (like the years $2019, 2020$), so Excel charted them as data. In **Select Data**, remove that series and use those cells as the axis labels instead.
- **A scatter plot shows both columns plotted against** $1, 2, 3, \ldots$ Excel made two series instead of one. In **Select Data**, remove both, click **Add**, and set **Series X values** to one column and **Series Y values** to the other. This is also how to swap which variable goes on which axis.
- **The chart is mostly empty space, or some bars are invisible.** Columns on very different scales were charted together (like price and volume), or the axis range was set by hand. Chart one scale at a time, and double click the vertical axis to check that its **Minimum** is 0 (**Reset** goes back to automatic).
- **There is no Histogram chart.** Older Excel (2013 and earlier on Windows, 2011 on a Mac) does not have one. Make the counts yourself: list the bin edges in a column, say `D2:D12`, then in `E2` enter `=COUNTIFS(A:A,">="&D2,A:A,"<"&D3)`, fill it down, and make a clustered column chart of the counts.
- **The trendline option is greyed out or missing.** Trendlines only work on some charts, such as scatter and column charts, and not on stacked or 3D ones. For a line of best fit, use a scatter chart.
- **The trendline equation has too many (or too few) digits.** Right click the equation, choose **Format Trendline Label**, and under **Number** pick **Number** with the decimal places you want.

A.5 Error Bars

- **Every bar has the same size error bar.** Excel is using a built in option (**Standard Error**, **Fixed Value**, or **Percentage**), or Specify Value points at a single cell. Open the Format pane, choose **Custom** $\rightarrow$ **Specify Value**, and highlight the whole column of $2 \cdot \text{SEM}$ values, one per bar.
- **Specify Value says there is a problem with the formula.** The old `={1}` was not cleared, so Excel combined it with the new cells. Delete everything in the box first, then highlight the cells. The box should end up looking like `=Sheet1!$H$2:$H$3`.
- **The error bars only go up (or only down).** Either **Direction** is set to **Plus** (or **Minus**), or the Negative Error Value was left as `={1}`. Set Direction to **Both** and fill in both boxes with the same cells.
- **The error bars are on the wrong bars.** The error bar cells must be in the same order as the bars. If you sort the summary table after making the chart, sort the whole table (means and error bars together).
- **The error bars are tiny or invisible.** Check the values in your error bar cells. A common slip is pointing at the standard deviation or count column instead of $2 \cdot \text{SEM}$. If the values are right, the error bars really are small compared to the axis, which is good news! Making them thicker under **Line** in the Format pane helps them show up.

- **Only one color of bars has error bars.** With more than one series, each series has its own error bars. Select the other series (click its bars, or use the dropdown at the top of the Format pane) and add them again with that series' column of values.
- **A scatter plot has horizontal error bars too.** Excel adds both directions to scatter plots. Click a horizontal error bar and press **Delete**.
- **I cannot find the + button (Mac).** Mac Excel does not have one. Use **Chart Design** → **Add Chart Element** → **Error Bars** → **More Error Bars Options**. If **Chart Design** is missing from the ribbon, click on the chart first.