Displaying Descriptive Statistics

- Displaying Quantitative Data
- Displaying Qualitative Data

} Single Variable

- Displaying Two Variables (*discuss in Chapter 3*)

- Reading: Chapter 2 (Sections 2.1 – 2.3)
  - Skip Polygons and Pareto Charts
- Optional Reading:
  - An Economist's Guide to Visualizing Data. Jonathan A. Schwabish. The Journal of Economic Perspectives, Vol. 28, No. 1 (Winter 2014), pp. 209-233 (posted on Canvas)

Displaying Quantitative Data

Recall the types of data: **Qualitative and Quantitative**

Summarizing Quantitative Data

- Frequency Distribution
- Relative Frequency Distribution
- Cumulative Relative Frequency Distribution
- Histogram

Tabular

Graph

Frequency Distribution

A **frequency distribution** shows the number of data observations that fall into specific intervals (classes)

Example: Number of iPads sold per day

Table 2.2 **The Number of iPads Sold in Each of 50 Days**

4	2	3	2	5
5	1	3	3	2
3	2	2	3	2
2	2	3	0	1
3	1	1	5	4
1	2	4	3	5
2	0	0	3	2
3	3	3	2	2
0	4	2	4	3
1	1	4	0	1

Table 2.3 **Frequency Distribution for the Number of iPads Sold in the Past 50 Days**

NUMBER SOLD PER DAY	FREQUENCY
0	5
1	8
2	14
3	13
4	6
5	4
Total	50

class

Relative Frequency Distributions

Relative frequency distribution displays the *proportion* of observations in each class *relative to the total number of observations*

- Shows the fraction of observations in each class
- Found by dividing each frequency by the total number of observations
- Fractions in a relative frequency distribution add up to 1

Relative Frequency Distributions

Example:

Table 2.4 **Relative Frequency Distribution for the Number of iPads Sold in the Past 50 Days**

NUMBER SOLD PER DAY	FREQUENCY	RELATIVE FREQUENCY
0	5	5/50 = 0.10
1	8	8/50 = 0.16
2	14	14/50 = 0.28
3	13	13/50 = 0.26
4	6	6/50 = 0.12
5	4	4/50 = 0.08
Total	**50**	**1.00**

Two iPads were sold on 28% of the days

Cumulative Relative Frequency Distributions

Cumulative relative frequency distribution totals the proportion of observations that fall *below the upper limit* of each class

- Shows the accumulated proportion as class values vary from low to high
- Cumulative relative frequency for the highest class is equal to 1

Cumulative Relative Frequency Distributions

Example:

Table 2.5 **Cumulative Relative Frequency Distribution for the Number of iPads Sold in the Past 50 Days**

NUMBER SOLD PER DAY	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
0	5	0.10	0.10
1	8	0.16	0.10 + 0.16 = 0.26
2	14	0.28	0.26 + 0.28 = 0.54
3	13	0.26	0.54 + 0.26 = 0.80
4	6	0.12	0.80 + 0.12 = 0.92
5	4	0.08	0.92 + 0.08 = 1.00
Total	**50**	**1.00**	

Three iPads or less were sold on 80% of the business days

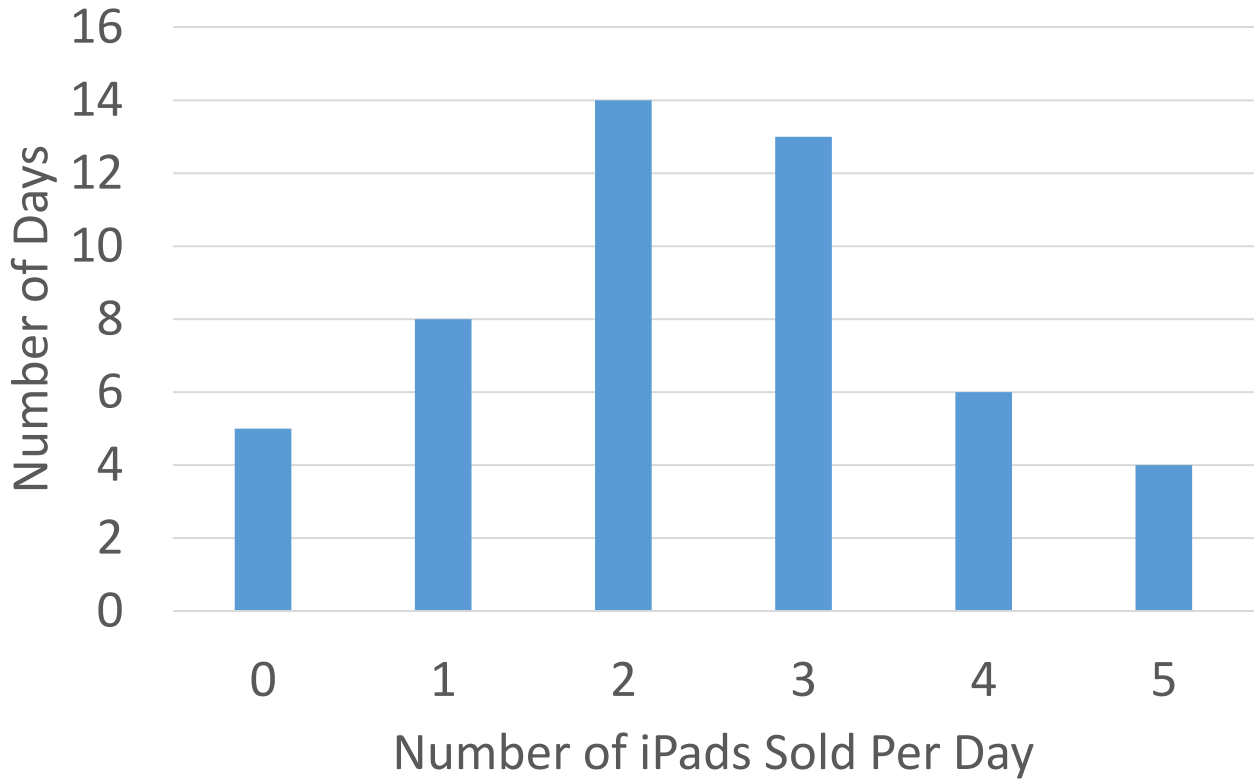
Histogram to Graph a Frequency Distribution

Histogram is a graph showing the number or % of observations in each class

- It is a graphical representation of a frequency distribution or the relative frequency distribution
- Classes of a variable of interest are placed on the horizontal axis
- A rectangle is drawn above each class interval with its height corresponding to the *frequency* or *relative frequency*

Histogram to Graph a Frequency Distribution

A **histogram** for the iPad example:



Excel Exercise >>

Discrete vs. Continuous Data

Discrete data are typically represented by integer numbers

- based on observations that can be counted (how many)
- take on whole numbers such as 0, 1, 2, 3

Continuous data are values that can take on any real numbers, including numbers that contain decimal points

- based on observations that can be measured (how much)
- take on any numbers such as 1, 3.1, 5.07, 4.941, etc.

Discrete vs. Continuous Data

Examples of Discrete data

- Number of children per family
- Number of cars listed per insurance policy
- Vacation days per month

Examples of Continuous data

- Time required to read Chapter 2
- Thickness of paint applied to a car body
- Person's height

Frequency Distribution Using Grouped Quantitative Data

Ideally, the number of classes in a frequency distribution should be between 4 and 20

- Some data sets, particularly those with continuous data, require several values to be grouped together in a single class
- This grouping prevents having too many classes in the frequency distribution, which can make it difficult to detect patterns

Class Width

There are methods to determine the number of classes k in a frequency distribution. But they are just a recommendation. You can always adjust!

Once k is known, the width of each class can be found as:

$$\text{Estimated class width} = \frac{\text{Maximum data value} - \text{Minimum data value}}{k}$$

- The width is the range of numbers to put into each class
- Round this estimate to a useful whole number that makes the frequency distribution more readable

Class Width

- There is no one correct answer for the class width
- The goal is to create a histogram to clearly and usefully show the pattern in the data
- Often there is more than one acceptable way to accomplish this

Class Boundaries

Class boundaries represent the minimum and maximum values for each class

Choose class boundaries that are easy to read:

☹☒

3.21 to less than 6.21 minutes
6.21 to less than 9.21 minutes

vs.

☺☑

3 to less than 6 minutes
6 to less than 9 minutes

Class Frequencies

Find class frequencies by counting and recording the number of observations in each class:

- Each class is represented by a range of values

Example:

Table 2.7 | **Frequency Distribution for Dell's Customer-Support Hold Times**

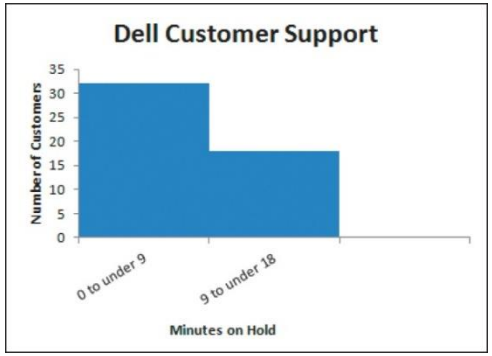
NUMBER OF MINUTES	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
0 to less than 3	5	0.10	0.10
3 to less than 6	18	0.36	0.46
6 to less than 9	9	0.18	0.64
9 to less than 12	11	0.22	0.86
12 to less than 15	6	0.12	0.98
15 to less than 18	1	0.02	1.00
Total	**50**	**1.00**	

Excel Exercise >>

The Consequences of Too Few/Many Classes

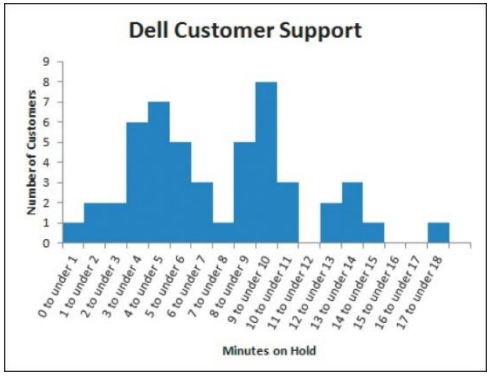
Wide classes results in few class intervals

- Can obscure important patterns
- Gives a "blocky" distribution graph
- Tells little about the distribution shape



Too many narrow classes in a histogram also has consequences

- Results in a "jagged" histogram
- Some classes may be empty



Displaying Qualitative Data

Qualitative data are values that are categorical

- Can be nominal or ordinal measurement level
- Describe a characteristic, such as gender or level of education

Summarizing Qualitative Data

- Frequency Distribution
- Relative Frequency Distribution

} Tabular

- Bar and Pie Charts → Graphs

Frequency Distributions

Frequency distribution:

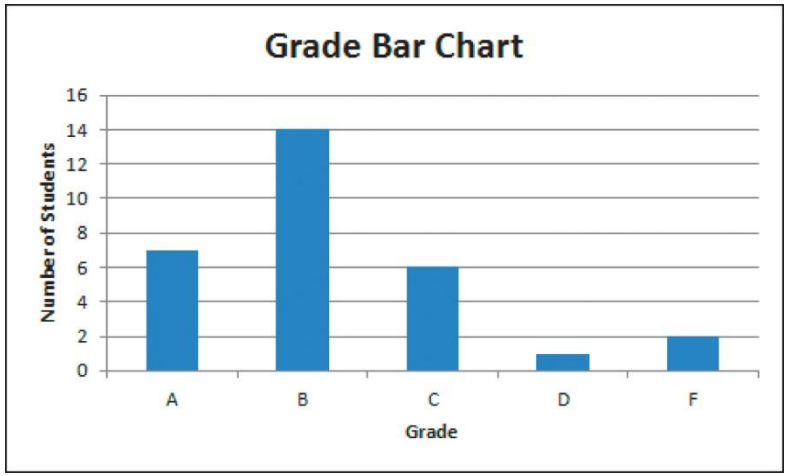
- Indicates the number of occurrences of various categories
- Techniques are similar to frequency distributions with quantitative data
- We can construct a relative frequency distribution (same idea as for the quantitative data)
  - Cumulative relative frequency distribution does not really make sense (specifically for nominal data)

Bar Charts

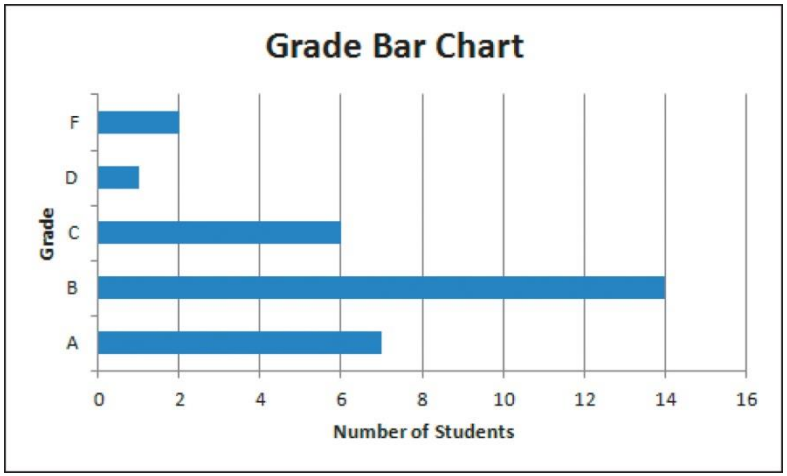
- Can be arranged in a vertical or horizontal orientation
- On one axis (usually, horizontal), we specify the labels that are used for each of the classes
- A frequency or relative frequency scale can be used for the other axis (usually, vertical)
- Using a bar of fixed width drawn above each class label, we extend the height appropriately

Bar Charts

Vertical bar chart



Horizontal bar chart



Excel Exercise >>

Pie Charts

Pie charts are a tool for *comparing proportions* for qualitative data

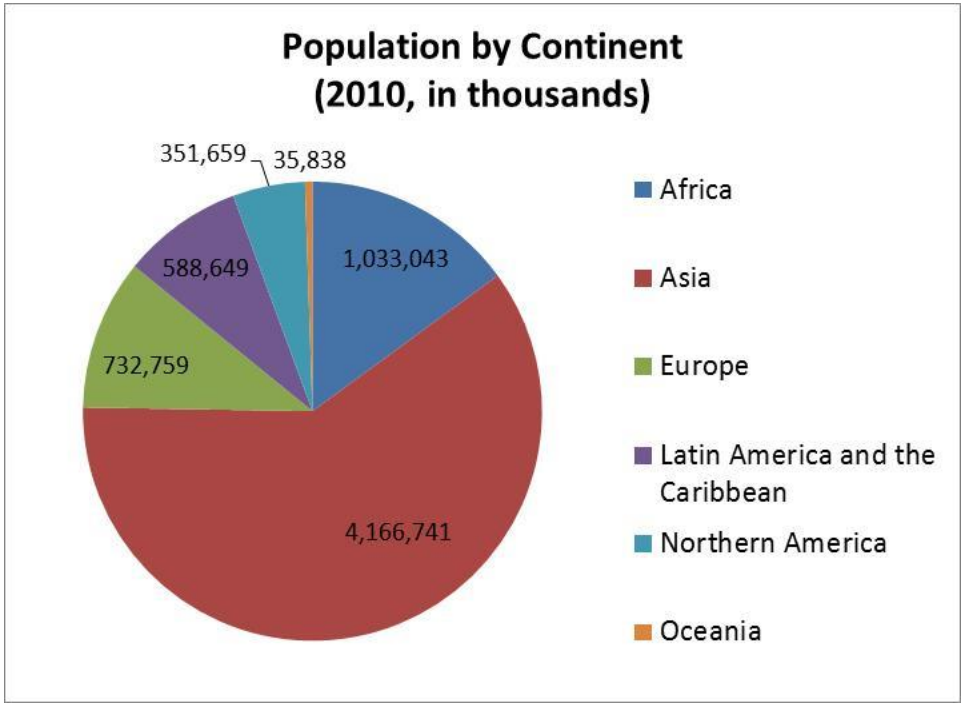
Each segment of the pie represents the relative frequency of one category

- All categories in the data set must be included in the pie
- Use a pie chart to compare the relative sizes of all possible categories
- Bar charts are more useful when you want to highlight the actual data values

Pie Charts

Example:

Continent	2010 population (in thousands)
Africa	1,033,043
Asia	4,166,741
Europe	732,759
Latin America and the Caribbean	588,649
Northern America	351,659
Oceania	35,838



Excel Exercise >>

Excel Time: Exercise 2.7

A major U.S. airline records the number of no-shows on a flight that operates each day from *Philadelphia* to *Paris*. A no-show is a passenger who purchases a ticket but fails to arrive at the gate at time of departure. The data for no-show during 70 flights can be found in the Excel file **no-shows.xlsx** (*Excel Files* folder → *Ch 02* on Canvas)

a. Construct a frequency distribution for these data.

b. Using the results from part a, calculate the relative frequencies for each class.

c. Using the results from parts a and b, calculate the cumulative relative frequencies for each class.

d. Construct a histogram for these data.

Back <<

Excel Time: Displaying Quantitative Data Using Excel

Frequency distributions indicate the number of observations that fall into a specific interval (class)

- First, define the bins in Excel (classes)
- Use Excel's **FREQUENCY()** function to count the number of values that occur within a range of values
  1. Highlight the cells where the frequency distribution will be placed
  2. With these cells highlighted, type =FREQUENCY() and select the array of data and the bins. **DO NOT HIT ENTER YET!**
  3. Now, hit simultaneously **Ctrl+Shift+Enter**

- *Note*: you won't be able to delete the result created by the FREQUENCY() function one cell at a time. If you need to correct something, delete all data and redo your calculations.

Excel Time: Displaying Quantitative Data Using Excel

Additional notes on the FREQUENCY function…

Simplifications for the current version of Office 365:

1. Highlight cells where the frequency distribution will be placed
2. With these cells highlighted, type =FREQUENCY() and select the array of data and the bins
3. **Hit Enter** (instead of Ctrl+Shift+Enter)

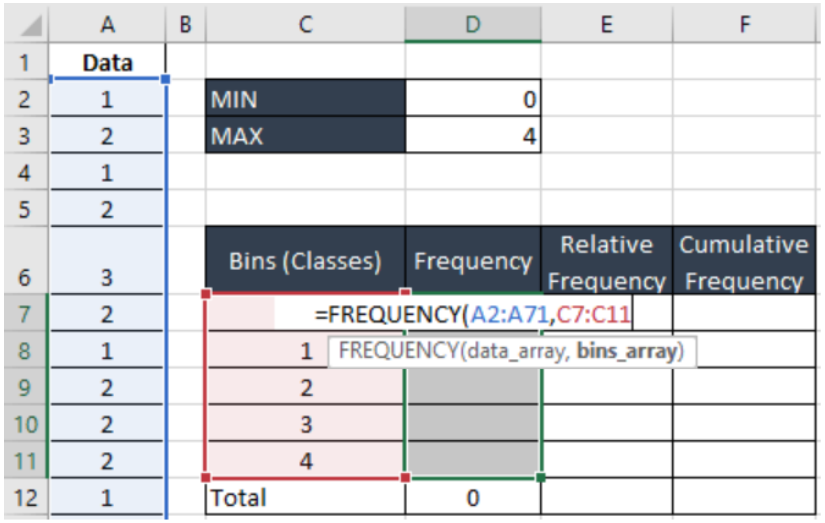
OR

1. Type the FREQUENCY function **only for the first bin/class** and select the array of data and ALL the bins
2. **Hit Enter**

Excel may add an extra element at the bottom in the frequency array (typically, zero). This element returns the count of any values *above the highest bin*. If Excel on your device adds an additional zero, just ignore it and proceed with the remaining calculations

Excel Time: Displaying Quantitative Data Using Excel

Frequency Distribution:

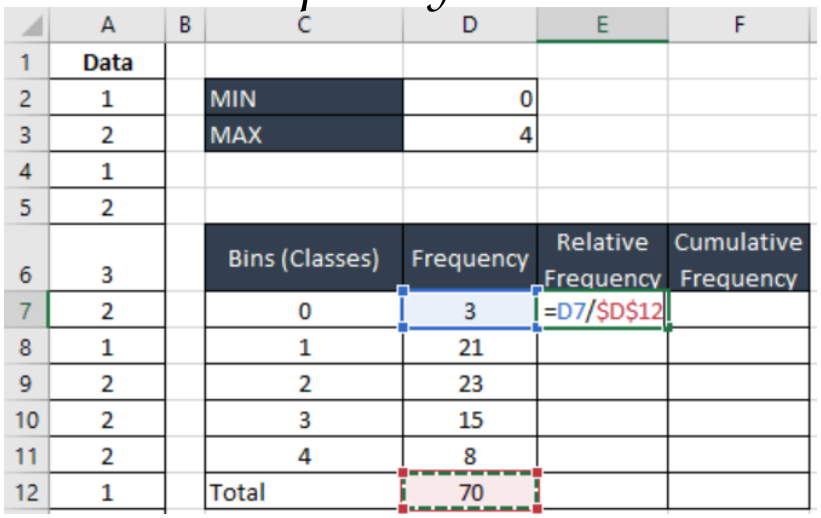


	A	B	C	D	E	F
1	Data					
2	1		MIN	0		
3	2		MAX	4		
4	1					
5	2					
6	3		Bins (Classes)	Frequency	Relative Frequency	Cumulative Frequency
7	2			=FREQUENCY(A2:A71,C7:C11		
8	1		1	FREQUENCY(data_array, **bins_array**)		
9	2		2			
10	2		3			
11	2		4			
12	1		Total	0		

Important: cells D7:D11 are highlighted when FREQUENCY function is entered

After that, hit Ctrl+Shift+Enter

Relative Frequency Distribution:



	A	B	C	D	E	F
1	Data					
2	1		MIN	0		
3	2		MAX	4		
4	1					
5	2					
6	3		Bins (Classes)	Frequency	Relative Frequency	Cumulative Frequency
7	2		0	3	=D7/$D$12	
8	1		1	21		
9	2		2	23		
10	2		3	15		
11	2		4	8		
12	1		Total	70		

Important: We divide frequencies by the total number of observations (same number!) ⇒ fix D12 in the formula

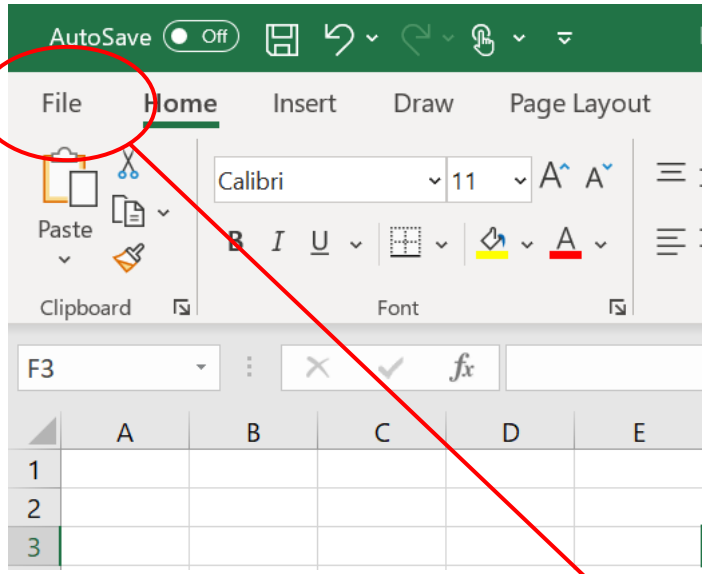
Adding a dollar sign seals cell's address ⇒ it won't change when you copy the formula down

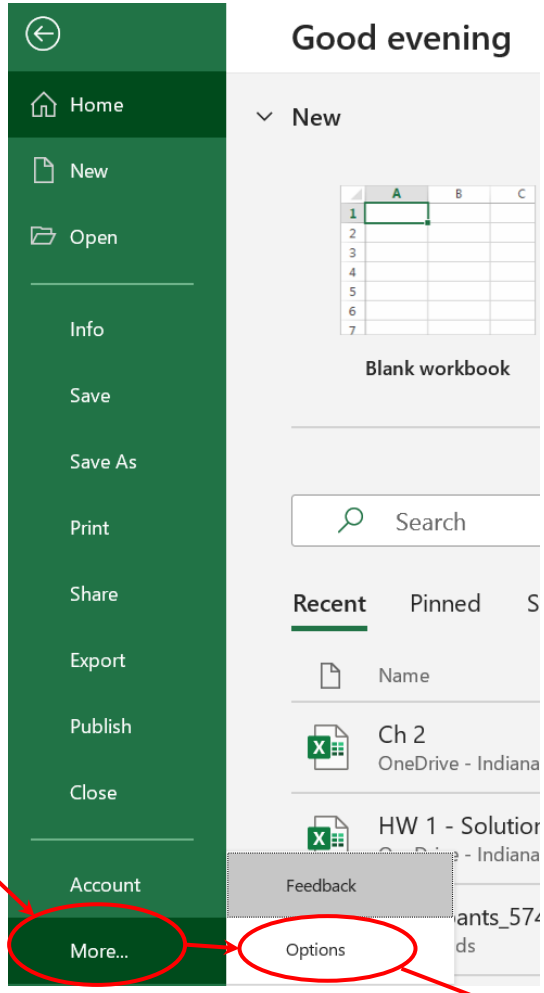
Excel Time: Displaying Quantitative Data Using Excel

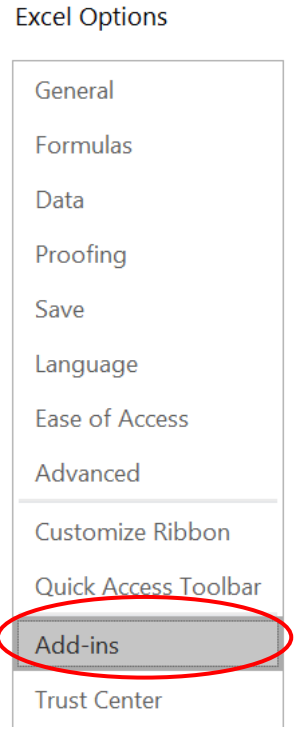
Excel offers two ways to construct a histogram:

- **If you have raw data**
  - *Data* > *Data Analysis* > *Histogram* > *OK*
  - *Note*: If you do not see the Data Analysis option under Data, you must add in this option: see the next three slides
- **If you have a frequency distribution or a relative frequency distribution**
  - Select the classes and the frequencies and use *Insert* > *Charts* > *Insert Column or Bar Chart* > *2-D Column* and choose the graph on the top left

Excel Time: Data Analysis Tool



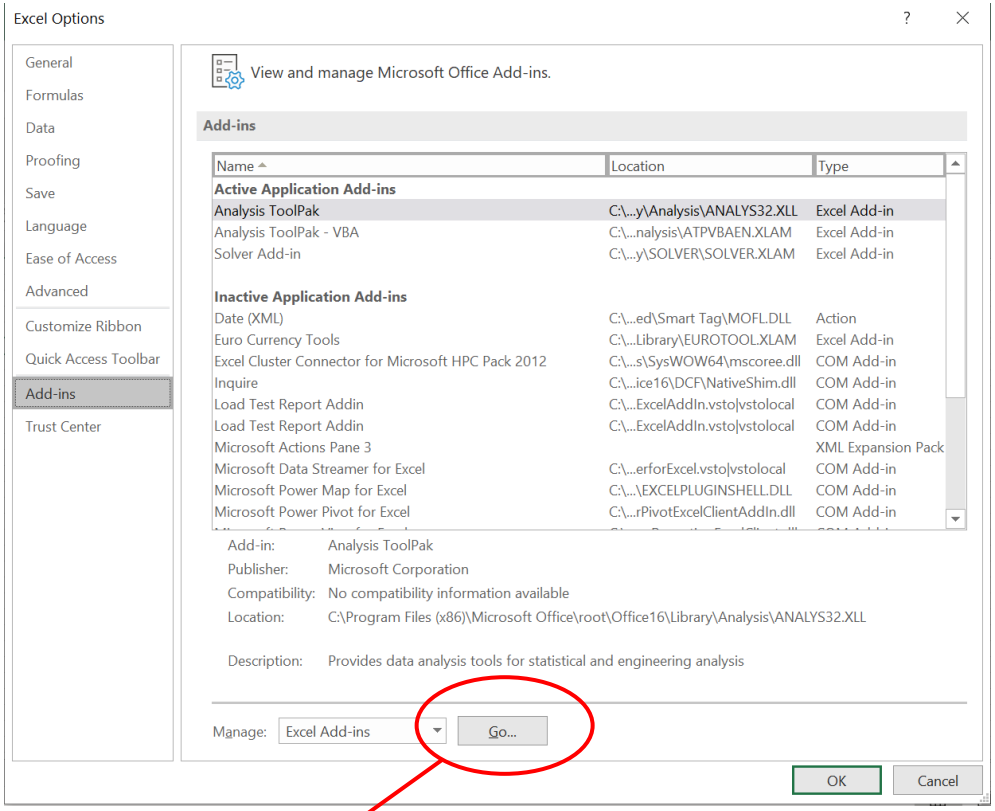




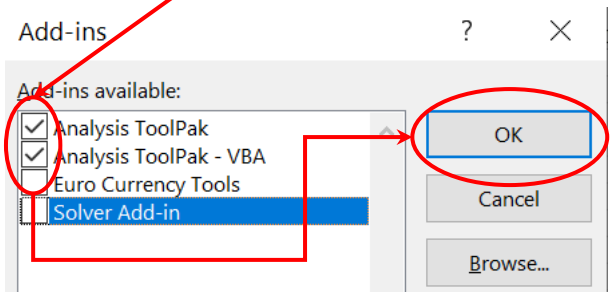
1. In Excel, click on the ***File*** tab
2. Click **Options** shown in the drop-down menu (may be hidden in the **More…** menu). This will open **Excel Options** dialog box
3. Select **Add-Ins** in the left margin of the **Excel Options** dialog box

Excel Time: Data Analysis Tool

4. Click on **Go…** at the bottom of the form



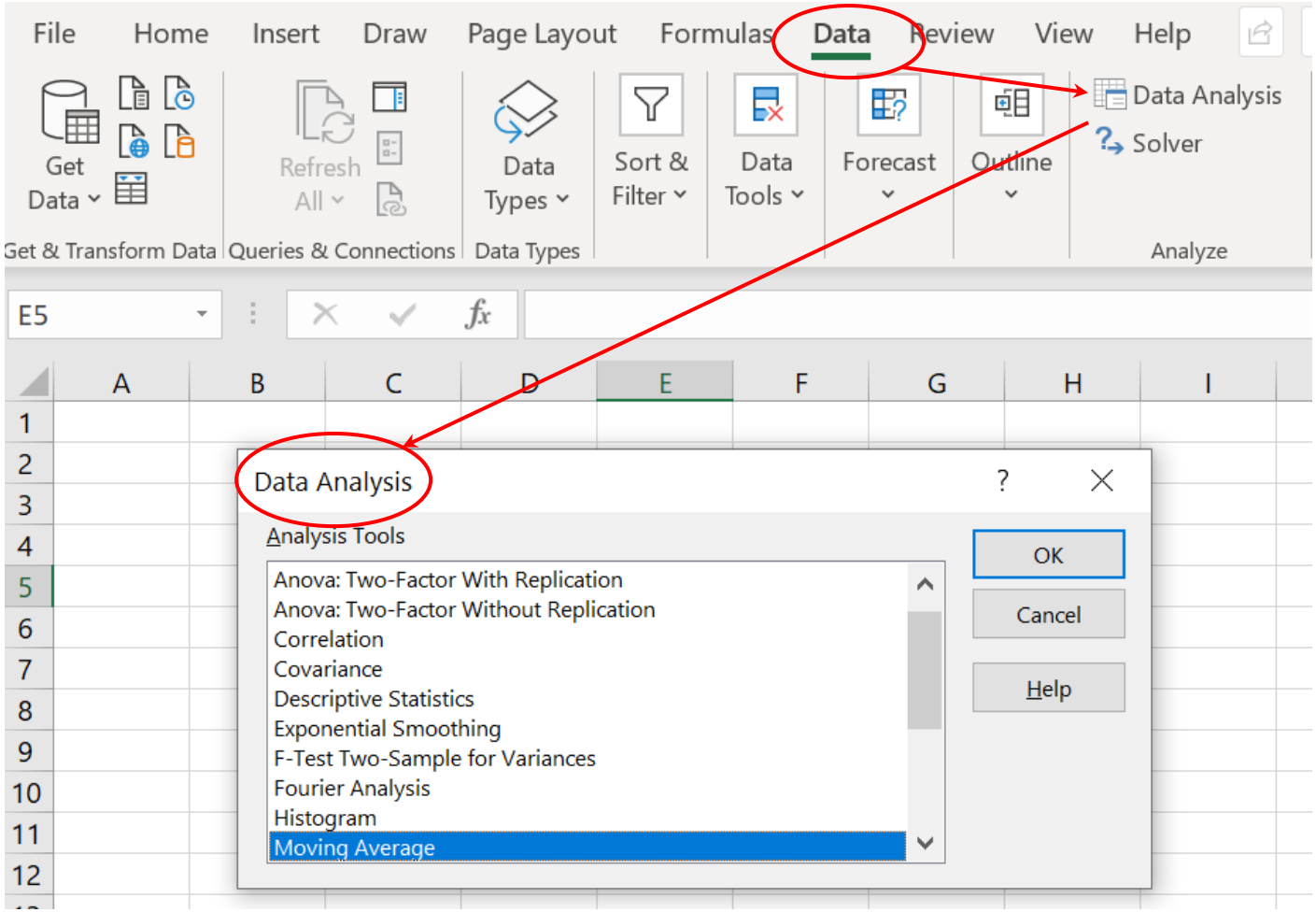
5. Check boxes for **Analysis ToolPak** and **Analysis ToolPak - VBA** in the popup menu and click **OK**



Excel Time: Data Analysis Tool

Select the *Data* tab. Click on **Data Analysis** on the right side of the application bar

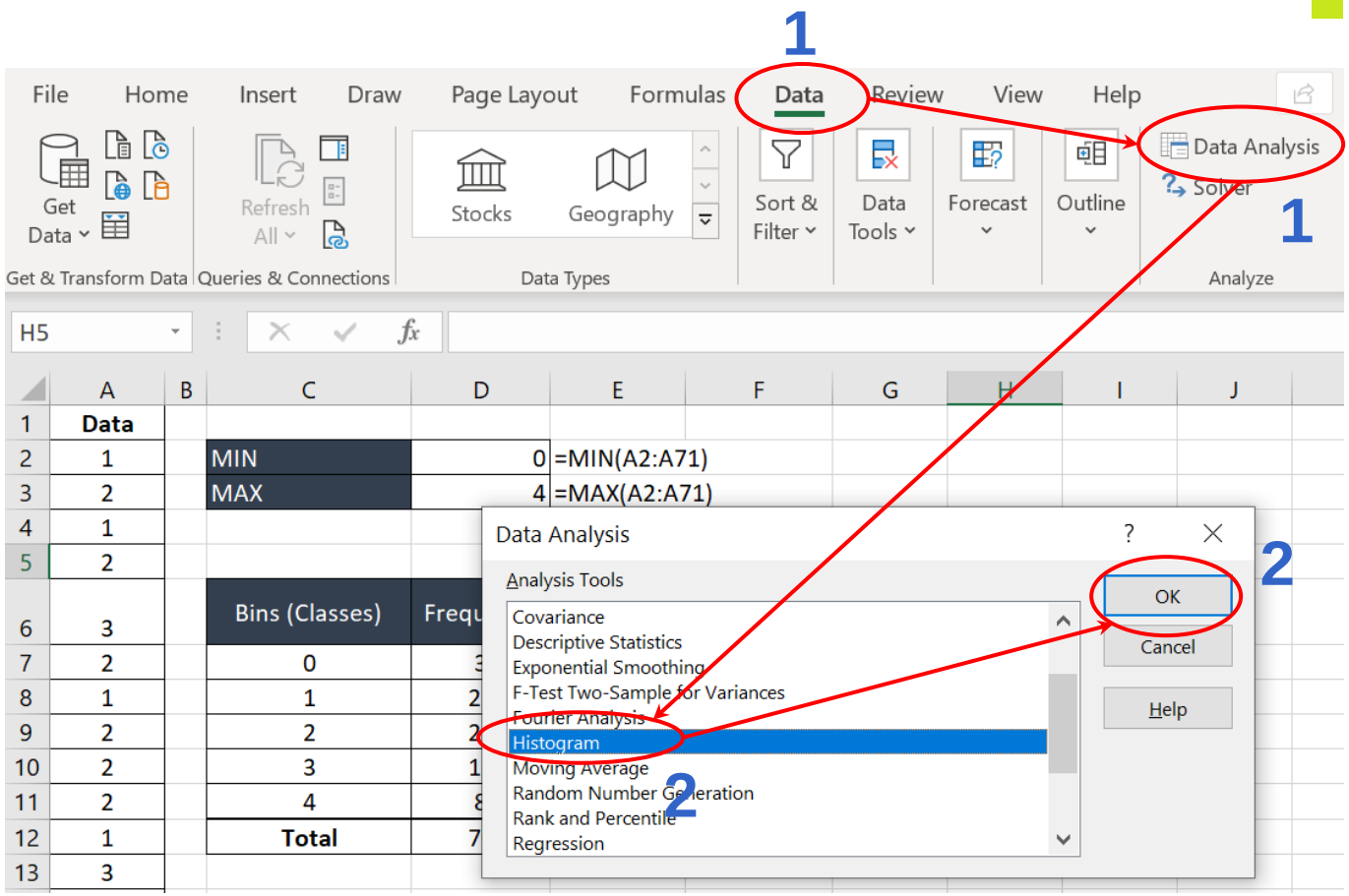
The Data Analysis pop-up menu should appear in the spreadsheet



Excel Time: Constructing a Histogram

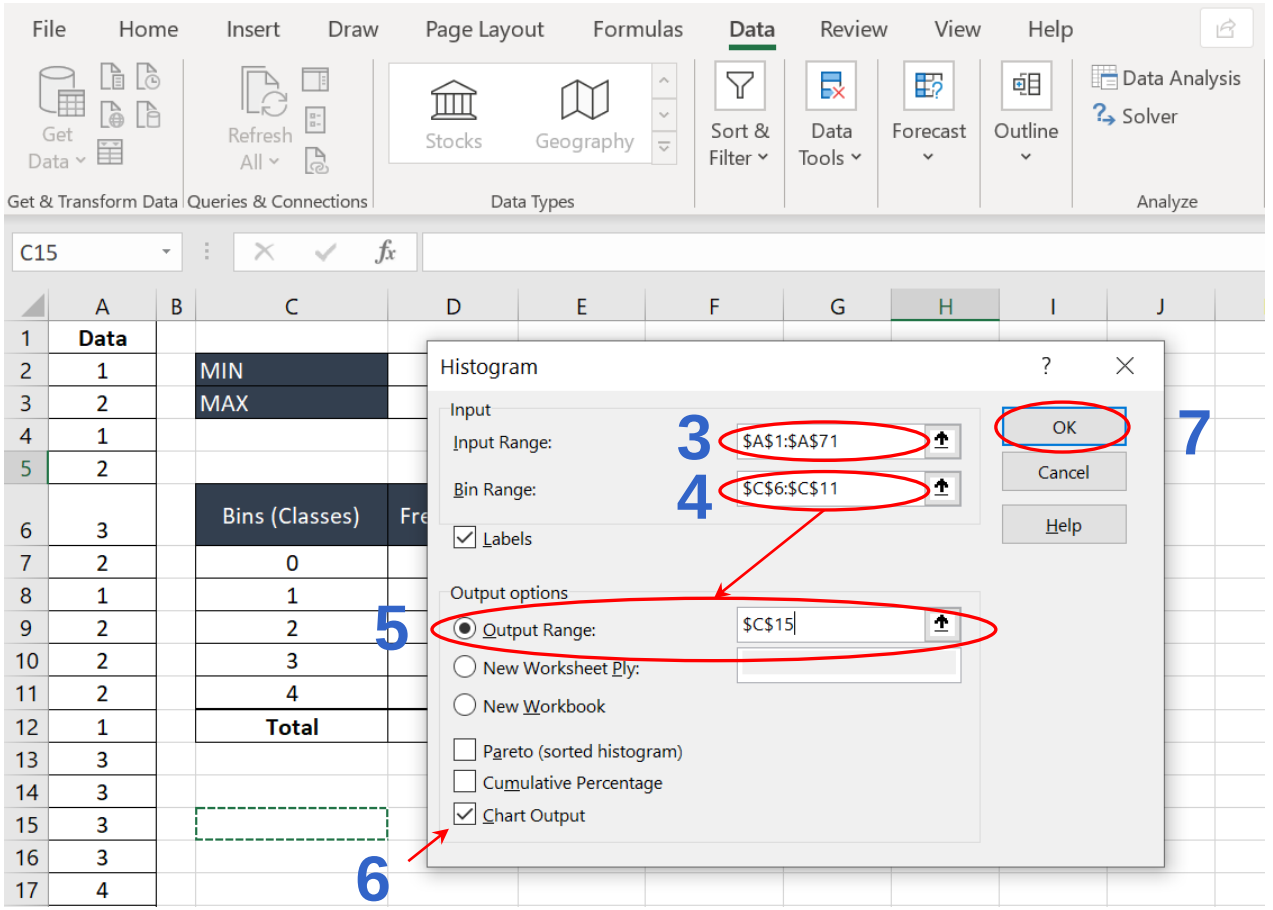
1. Select the **Data** tab, and click on **Data Analysis** in the upper right corner

2. In the pop-up menu, select **Histogram** and click **OK**…



Excel Time: Constructing a Histogram

3. In the **Input Range**, highlight the data
   - Check «Labels» if you selected the data with the column name
4. In the **Bin Range**, highlight the bins (create if not already created before step 1)
5. For **Output options**, select where you want to see the results
   - Specifying one cell indicates the upper left corner of the output
6. Choose "Chart Output" option if you want the histogram to be displayed
7. Click **OK**



Excel Time: Exercise 2.5

Excel file **college_credit_card.xlsx** (*Excel Files* folder → *Ch 02* on Canvas) contains the results of a survey that collected the current credit card balances for 36 undergraduate college students.

a. Construct a frequency distribution for these data.

b. Using the results from part a, calculate the relative frequencies for each class.

c. Using the results from parts a and b, calculate the cumulative relative frequencies for each class.

d. Construct a histogram for these data.

Back <<

Excel Time: Bins for Grouped Data

- For grouped data, it is important to create the bins correctly
- In Excel, the bins represent the **upper bound** of the class
- For example, if the class is represented by the values between 0 and 10, we actually mean that values ≥ 0 but **< 10** fall in this class
  - → We want to define a bin such that data points equal to exactly 10 DO NOT fall in this class
  - → We can define a bin to be just a little bit lower than 10!!! For example, 9.9999 (or 9.99).
  - → *A word of caution*: when decreasing the upper bound, make sure that no data below 10 is forced into the next class
    - → The best way to do that is to reduce the bin by a very small number (e.g. 0.0001, maybe less depending on the data)

Excel Time: Exercise 2.12

A restaurant manager surveys her customers after their dining experience. Customers rate their experience as *Excellent* (E), *Good* (G), *Fair* (F), or *Poor* (P). Experience of 60 customers is recorded in the Excel file **dining_experience.xlsx** (*Excel Files* folder → *Ch 02* on Canvas).

a. Construct a frequency distribution for these data.

b. Using the results from part a, calculate the relative frequencies for each class.

c. Construct a vertical bar chart for these data.

d. What percentage of customers rated their dining experience as either Excellent or Good?

Back <<

Excel Time: Displaying Qualitative Data

Frequency distributions display qualitative data by indicating the number of occurrences of various categories

- Use Excel's **COUNTIF()** function to count the number of values matching a category label

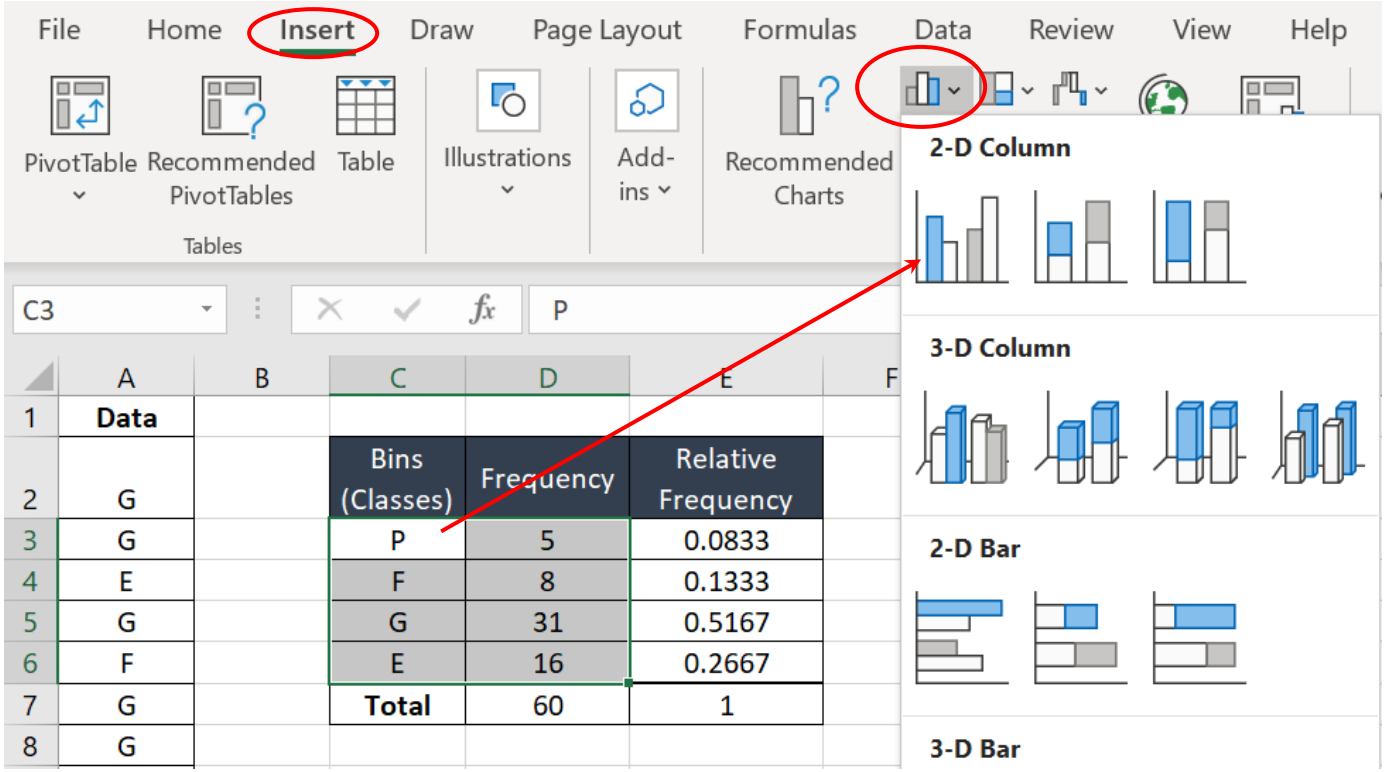
	A	B	C	D	E
1	**Data**				
2	G		Bins (Classes)	Frequency	Relative Frequency
3	G		P	=COUNTIF($A$2:$A$61,C3)	=D3/$D$7
4	E		F	=COUNTIF($A$2:$A$61,C4)	=D4/$D$7
5	G		G	=COUNTIF($A$2:$A$61,C5)	=D5/$D$7
6	F		E	=COUNTIF($A$2:$A$61,C6)	=D6/$D$7
7	G		**Total**	=SUM(D3:D6)	=SUM(E3:E6)

- *Useful Notes*:
  - Fix the array of data inside COUNTIF function: when you copy the formula in D3 down, the array address won't change
  - D7 is fixed for calculating relative frequencies
  - Hitting F4 on the keyboard adds the dollar signs

Excel Time: Bar Charts

Bar chart is a tool for displaying *qualitative* data that have been organized in categories

Can be arranged in a vertical or horizontal orientation



Excel Time: Exercise 2.17

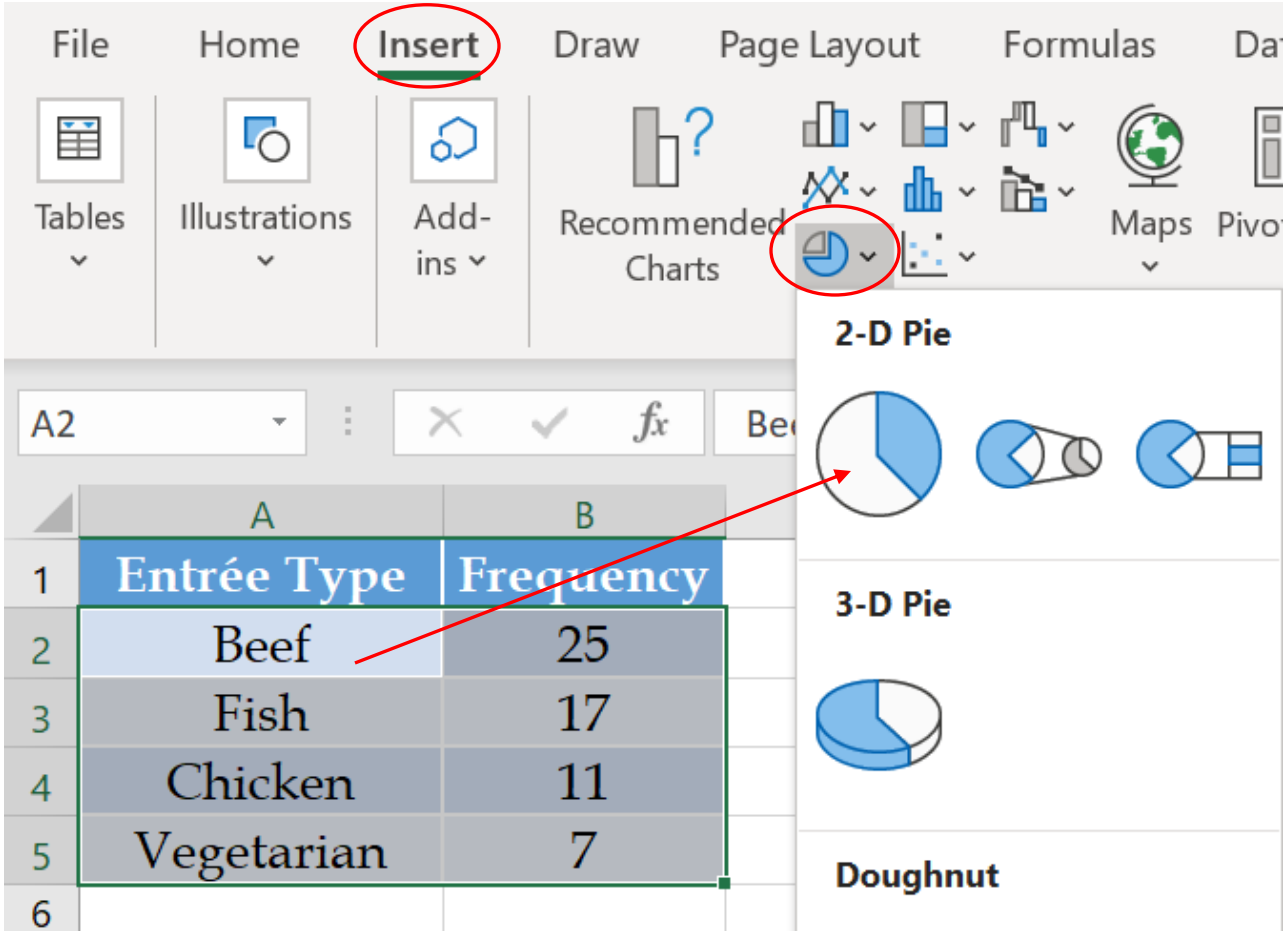
A manager at a local restaurant records the number of customers who ordered each of four types of entrees. The following table summarizes the results:

Entrée Type	Frequency
Beef	25
Fish	17
Chicken	11
Vegetarian	7

Construct a pie chart summarizing these data.

Back <<

Excel Time: Pie Chart



Exercise 2.9 (Extra Practice)

The Excel file labeled **Lowes.xlsx** (*Excel Files* folder → *Ch 02* on Canvas) lists the receipt total for 350 randomly selected customers for *Lowes*, a home improvement store.

a. Construct a frequency distribution for these data.

b. Using the results from part a, calculate the relative frequencies for each class.

c. Calculate the cumulative relative frequencies for each class.

d. Construct a histogram for these data.

Exercise 2.10 (Extra Practice)

The Hawaii Island Chamber of Commerce collects ocean temperature data to help promote tourism for local businesses. Excel file titled **Hawaii_ocean_temps.xlsx** (*Excel Files* folder → *Ch 02*) lists daily ocean temperatures for the past 125 days.

a. Construct a frequency distribution for these data.

b. Using the results from part a, calculate the relative frequencies for each class.

c. Using the results from part a and b, calculate the cumulative relative frequencies for each class.

d. Construct a histogram for these data.

e. For what percentage of days was the ocean temperature less than 78 degrees Fahrenheit?

Exercise 2.11 (Extra Practice)

Excel file **search_engine.xlsx** contains the results of a survey that asked users of the Internet to identify their favorite search engine (FYI, Baidu is a Chinese-language search engine ☺)

a. Construct a frequency distribution for these data.

b. Using the results from part a, calculate the relative frequencies for each class.

c. Construct a vertical bar chart for these data.

Exercise 2.18 (Extra Practice)

The following table shows the breakdown of gasoline prices in terms of its components:

Component	Percentage
Crude Oil	60.9
Taxes	15.1
Refining	13.7
Distribution and Marketing	10.3

Construct a display that best describes these data.