

City University of New York (CUNY)

## CUNY Academic Works

---

Open Educational Resources

LaGuardia Community College

---

2021

### Lab Exercises for Statistics Using Excel

Julia Nebia

*CUNY LaGuardia Community College*

Steven Cosares

*CUNY LaGuardia Community College*

Milena Cuellar

*CUNY LaGuardia Community College*

[How does access to this work benefit you? Let us know!](#)

More information about this work at: [https://academicworks.cuny.edu/lg\\_oers/103](https://academicworks.cuny.edu/lg_oers/103)

Discover additional works at: <https://academicworks.cuny.edu>

---

This work is made publicly available by the City University of New York (CUNY).

Contact: [AcademicWorks@cuny.edu](mailto:AcademicWorks@cuny.edu)



# Lab Exercises for Statistics using Excel

Julia Nebia  
Steven Cosares  
Milena Cuellar  
CUNY - LaGuardia Community College



Nebia, Cosares, Cuellar, 2021. This work is licensed under a Creative Commons Attribution-Non Commercial-Share Alike 4.0 International License.

## A Note to Instructors

This document contains the text associated with a series of computer-based lab exercises to help students apply the concepts usually included in a first course in Statistics. A compressed file has been included that contains a separate folder for each lab. In each folder is an excel spreadsheet file and an editable word document providing the instructions for students to complete the exercise. The exercises are not numbered in the folders, so you can select any subset of these exercises to assign to your students. Consistent with the Creative Commons copyright associated with these exercises, you are free to modify the instructions in any way you see fit, e.g., to add activities or to articulate the format of the lab-reports that will be completed by the student or the student groups as part of the exercise. However, we request that the attribution at the end of the document remain intact.

## Exploring Excel Features

### Terminology

An Excel spreadsheet file organizes data as a ledger, which is a book that has been used by accountants and bookkeepers for years. When you open the file, you see a “workbook” that contains one or more tabbed “pages”. Each page contains many rows and columns of boxes called “cells”. Lists of (related) values can be placed down a column of cells, or across a row of cells, or in a table.

Open the excel file “Excel Practice” and select the page named “Small Table”. This page contains a table listing some data about different models of SUV. Notice that a cell can contain different kinds of values. You can put in a number that can represent a dollar amount, a count, a percentage, a date, etc. You can instead put in some text like letters, labels, words or sentences.

The value in each cell in a sheet can be “referenced” by its location. The columns are lettered, and the rows are numbered. The upper-left-most cell is in location A1.

1. What value is in cell D9? What value is in cell B2?

In a cell, you can also place “derived data” which is a value or piece of text that represents the result of some formula or other manipulation of the values in other cells. The type of value in the cell tells excel which operations are valid. For instance, you are allowed to add two numbers, but you cannot add two words.

2. Click on cell G5. Place the following “formula” into cell G5: “=C4 + C5”. Now G5 contains the sum of the values.
3. Click on cell C4 and type “0”, then press Enter. Notice that value in C4 is changed and the value in G5 reflects that change.

Excel recognizes formulas involving: ‘+’ (plus), ‘-’ (minus), ‘\*’ (times), ‘/’ (divide by), and ‘^’ (to the power of).

4. In cell G7 type the formula: “=3\*(5-2)^4/12-15” and press Enter. Notice that Excel respects Algebra’s rules of operation precedence.

Excel also provides some named “functions” to perform more complex operations. Through practice you will learn about many of these.

5. In cell G9 type: “=Max(C2,C3,C4,C5,C6)” and press Enter. Observe that the maximum value appears.
6. In cell G10 place the formula: =Max(C2:C6) (You should get the same result).

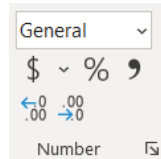
Functions that involve multiple values can take a “range” of cells as an argument. A range is defined as follows:      Upper-left location : Lower-right location

7. In cell G12 place: = Sum(C2:C21)
8. In cell G14 place: =Countif(D2:D21,”>70”)      Can you guess what this function does?

## Calculations

Along the top of the Excel screen is a list of “ribbons” which are menus for different operations on a spreadsheet. The “Home” ribbon has the features that are used most often.

1. Click on the tab on the bottom for the sheet labeled “Calculations”. The sheet contains a table representing sales for two consecutive years. In cell D2 you must place a formula that gives the percent change from 2020 (B2) to 2021 (C2). This is the difference divided by the earlier year’s value. (Notice that this value should be -13.9%). If the value in the cell is not expressed as a percentage, click on the “%” in the Number-Format submenu of the Home ribbon:





You may want to click on the icon on the lower left that increases the number of decimal places displayed for this value.

2. Click on cell D2 and Copy it. This is done by keying “Control-C” or right-clicking the mouse and selecting “Copy”. Click on cell D3 and Paste. This is done by keying “Control-V” or right-clicking the mouse and selecting “Paste” with the first option. Notice that the value in D3 depends on B3 and C3, i.e., that you copied and pasted the formula, not the contents.
3. Make sure to place the correct formulas into cells D4 and D5.
4. For cells B7 and C7 you can use the Sum( ) function to obtain the appropriate values.
5. Make sure you place the correct formula into cell D7. (The value should be -10.3%)

## Formatting

Click on the tab for the sheet labeled “Final”. The sheet has a table that is formatted. Use this sheet as a reference for the next set of tasks. The sheet “Formatting” contains the data in its original form. You are to do the formatting operations to make a chart that looks like the one in “Final”. For example:

1. The title is located in cell A1. Click and drag to cover cells A1:F1. Find the icon in the Home ribbon that allows you to merge and center the total over the whole table: . Increase the size of the font of the title and make it bold.
2. Use the Number-Format submenu to present the values in the table as dollar values. You don’t need to present the digit for the cents.
3. To place the lines in the table you can select from the drop-down menu that looks like .

**Extra Credit:** Use the “Insert” ribbon to create a Pie Chart that represents the data from the 2020 sales column. Format it to look as close as possible to the chart in the sheet “Final”.

## Random Samples

### PART I

**Instructions:** Open the spreadsheet file labeled “Random Samples”. Click on the web address located in cell A1 in sheet “Select”. Use the picture on the page for this activity. You will see 60 circles. This is the “population.” *Our goal is to estimate the average diameter of these 60 circles by choosing a representative sample.*

1. Click on any five circles on the page. Try to select a mix of circles with sizes that match the overall mix of the 60 circles on the page. Your selected circles will turn orange. Type the average diameter for these five circles into cell A3. (Make sure you have five orange circles before you record the average diameter.)
2. “Reset” the picture.
3. Choose another mix of five circles and record the average diameter for this second sample of circles. You can reuse a circle, but the sample should not have all the same circles as before. Type the average diameter for these five circles into cell A4. You now have the average values for two different samples.
4. Reset and repeat for a total of 10 different samples whose averages are located in cells A3 to A12.
5. Using excel, calculate the average of the 10 values you obtained by typing the following into cell A13:

“= average(A3:A12)”

6. Answer the following questions:

The average diameter for this population of 60 circles is 19.3. How many of your samples had an average diameter greater than 19.3? How many of your samples had an average diameter less than 19.3? How close is the “average of averages” to the population average of 19.3? If they are far apart, can you explain why?

### PART II

**Question:** Did you select representative samples in Part I?

Humans often think they are being unbiased in their selections when they are asked to be. Is it possible that you selected more circles of a certain size? Many statisticians believe that the best way to obtain a representative sample is to use tools that assure randomness in a selection, so that every member of the population has an equal likelihood of being in a sample.

7. Suppose you wanted to select circles by closing your eyes and moving the mouse to random spots on the screen. Explain why this approach may **not** generate a representative sample. (Hint: Look up the term “batch size bias”).

**Instructions:** Click on the web address located in cell A20 in sheet “Select”. You will again see the same 60 circles. As before, this is the “population.” *Our goal is to estimate the average diameter of these 60 circles by choosing a random sample.*

8. Click on the “Generate sample” button to get a random sample of five circles by clicking on the random sample button. The simulation randomly chooses five circles. Record the average diameter for the random sample into cell A22.
9. Reset the simulation using the “Reset” button.
10. Repeat the following nine more times to place sample averages into cells A22 to A31: Click on the “Generate sample” button. Record the average diameter for this random sample. Click “Reset”.
11. Using excel, calculate the average of the 10 values you obtained.
12. Answer the following questions:

The average diameter for this population of 60 circles is 19.3. How many of your samples had an average diameter greater than 19.3? How many of your samples had an average diameter less than 19.3? How close is the “average of averages” you calculated to the population average of 19.3?

Did the process you used in Part II provide more trustworthy estimates than the process in Part I?

## Descriptive Statistics in Excel

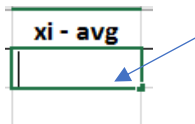
### Mean and Standard Deviation

In class, we have learned how to calculate the mean and standard deviation by hand for a small data set. The formula for the sample mean of  $x_1 \dots x_n$  is  $(\sum x) / n$ . The formula for the sample standard deviation is a bit more complicated:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

We apply these formulas by placing the data value in a chart. A copy of this chart is located in the sheet “Standard Deviation” in the Excel file “Descriptive Statistics”.

1. A sample data set with  $n=8$  is included in the sheet. Place the mean into cell B11 by using the formula “=(B2+B3+B4+B5+B6+B7+B8+B9)/8”. (Notice that a more compact formula is located in cell B15)
2. Place “=B2-B\$11” into cell c2. (The “\$” helps us copy the formula to other cells).
3. Copy the formula into cell C3...C9 by clicking on C2, grabbing the small square at the lower right corner and dragging down.



4. In cell D2 place the formula: “=C2^2”.
5. In cell D11 place the formula “=SUM(D2:D9)”.
6. In cell D12 type “=D11/(A9 – 1)”. This is the sample variance.
7. In cell D13 the sample standard deviation appears by using the formula: “=sqrt(D12)”. (Notice that a more compact formula is located in cell D15)
8. Below are salaries of thirteen players on the Cleveland Cavaliers basketball team during the 2009–2010 season.


\$736,000	\$9,300,000
\$6,364,000	\$4,089,000
\$1,429,000	\$4,254,000
\$736,000	\$2,644,000
\$855,000	\$458,000
\$21,000,000	\$3,000,000
\$11,541,000	

Use Excel to compute the mean and standard deviation of these sample values.



## Five-Number Summary

The Five-Number Summary describes the data by finding individual values that have some meaning. As a result, the statistics are not found through formulas, but through (search and sort) procedures.

1. In the sheet “Five Number Summary” a data set containing 200 values is located in the range [A1:A200]. To find the five values at key locations, we need to sort the values. Select all 200 data values and click on the ascending-sort button  on the “Data” ribbon.

2. The minimum is now in location 1, so place “=A1” into cell D4.

3. The maximum is in location 1, so place “=A200” into cell D8.

4. The median is at location  $(20 + 1)/2 = 10.5$ . Since this is not a whole number, the median is between the values in location 10 and location 11. So, place the following into cell D6:

$$\text{“}=(A10 + A11)/2\text{”}$$

5. The first quartile is in location  $(20 + 1)/4 = 5.25$ . Since this is not a whole number, the 1Q value is between the values in location 5 and location 6. So, you can split the difference by placing the following into cell D5:

$$\text{“}=(A5 + A6)/2\text{”}$$

6. The third quartile is in location  $3*(20 + 1)/4 = 15.75$ . Since this is not a whole number, the 3Q value is between the values in location 15 and location 16. So, you can split the difference by placing the following into cell D5:

$$\text{“}=(A15 + A16)/2\text{”}$$

7. To find a measure of “spread” and to identify if any values qualify as “outliers”, we calculate the Interquartile Range (IQR). Place “=(D8-D5)” into cell G6.

8. The Lower Fence = Median – 1.5 IQR. Place that value into cell G5.

9. The Upper Fence = Median + 1.5 IQR. Place that value into cell G7.

10. Does the data contain any outliers?

11. Excel provide functions for these statistics that do not require sorting the data.

Place “=MINIMUM(J1:J200)” into cell M4.

Place “=QUARTILE(J1:J200,1)” into cell M5.

Place “=MEDIAN(J1:J200)” into cell M6.

Place “=QUARTILE(J1:J200,3)” into cell M7.



Place “=MAXIMUM(J1:J200)” into cell M8.

12. Can you explain why it is OK if there is a slight difference between your quartile values and those created by excel?

## Visual Descriptions of Data


### How to create a Dot Plot

Recall that a Dot-Plot represents each value in a data set by a dot placed on the number line, we stack any repeated values to give shape to our description. Open the Excel Spreadsheet file called “Visual Descriptions” and select the sheet “Dot Plot”.

1. Suppose you had a data set like the list of values in range [A1:A25].
2. Select the data located in the range. Sort the data into ascending order by selecting the “Data” ribbon and clicking the “Sort” Icon. Chose ascending order:  .
3. Type the number 1 into cell B1, next to the smallest data value.
4. In cell B2 type the following formula:  
`=IF(A2=A1, B1+1,1)`  
This formula will assign increasing values when a data item is repeated.
5. Copy the formula to cells B3 to B25 by moving the cursor to the bottom right corner of the cell until the points becomes a black plus +. Then click and drag down the length of the data list, e.g., from B2 to B25.
6. Select both columns from A1:B25, go to the “Insert” top menu, and select the Scatterplot  icon.
7. Can you explain why this type of picture makes the most sense when the data values come from a small range of whole numbers?

### How to create a Pie Chart

Pie charts are used to describe the breakdown of a population into its important subgroups. Go to the sheet labeled “Pie Chart”. The size of the pie is in direct proportion to the size of the subgroup.

1. The data contained in the range A1:E20 lists the party affiliations of 120 people. There are 6 different parties represented. They are listed in column I.
2. In cell J2 enter the number of times that “Working Families” is listed, this is the “Frequency”.
3. In cell J3 type the following: `=COUNTIF(A$1:E$20,I3)`
4. Copy this COUNTIF() function into the remaining 4 rows in column J. Confirm that the sum of the counts is equal to 120.
5. Select the categories and counts in the range I1:J7
6. Go to the “Insert” top menu and click on the Pie Chart icon  and select the 2D pie chart. Which party has the largest number in the sample?
7. Excel provides many options to allow you to format the chart you created. Try some of these. Can you display the relative frequency (percent) for each piece of the pie?

8. Can you explain why this type of picture is better suited for categorical data than numerical data?

### **How to create a Histogram**

When number data is partitioned over equal length segments on the number line, it makes more sense to place the shapes representing the size of each group on the number line than to use a pie chart. So instead of being represented as pieces of a pie, each group appears as a rectangle, whose height is proportional to the number of data points in the segment. The total of the heights should equal 100%. Excel has add-in features to draw histogram, but we have better control of the result when we construct our own. Go to the sheet labeled “Number Data”.

1. The 200 data values are located in the range A2:A201. The range of the data values is a minimum of 43 and a maximum of 118. The page is missing the mean of the values. Type “=AVERAGE(A2:A201)” into cell D3.
2. The page is also missing the standard deviation of the values. Type “=STDEV(A2:A201)” into cell D4.
3. Two of the bars are missing in the histogram. That’s because the associated rows in the Frequency Chart are missing. Count the number of values in the data that are in the range [54, 66] and place the count into cell D15.
4. In cell E15 place the formula for the percentage “=D15/D20”.
5. Excel can count for you. Place the following into cell D16: “=COUNTIF(A\$2:A\$201,“<=79”) - COUNTIF(A\$2:A\$201,“<=66)”.
6. In cell E16 place the formula for the percentage “=D16/D20”. Make sure that all 200 values are accounted for.
7. Can you explain why the first and the last rectangles are slightly shorter than they should be. (Hint: compare the range of the data with the range of the number line in the picture).

## The Law of Large Numbers

The field of Statistics supports our desired to learn about a population, based on the information provided from some sample. In essence, the law of large numbers says that you can get more reliable information from larger samples than from smaller ones, presumably because larger samples are more likely to be representative of the complete picture associated with the population, in all of its diversity and possibility.

Suppose we have a coin, but we are not sure whether it is “fair” or if it has a bias. In other words, we want to identify the value of  $p$ , the probability of showing a Head when it is flipped. We can think of  $p$  as a population parameter, i.e., that if we observed the outcome from all (infinite number of) flips the coin could take, then  $p = (\#Heads / \#Flips)$ .

Obviously, in any statistical study of the coin, we can only perform a finite number of flips. We might be interested in how closely we can approximate  $p$  from such a sample. Open the spreadsheet file “Biased Coin Flips”.

1. Go to the “Single Flips” sheet. Click on the button “New Game”. This creates a brand-new coin with some  $p = 0.1$  or  $0.2$  or  $0.3 \dots$  or  $0.9$ . You have to guess the value for  $p$ .
2. Press “Flip Coin” 5 times. Based on this experience, a value shows in “Head Pct”, which =  $(\#Heads / \#Flips)$ . Do you think this value is close to  $p$ ? Do you think you could conclude whether  $p=0.5$  or if it is greater or smaller?
3. Press “Flip Coin” 5 more times. Based on this experience, a value shows in “Head Pct”. Do you think this value is closer to  $p$  than before? Do you have more confidence in your guess about whether  $p=0.5$  or if it is greater or smaller?
4. Make a guess about the value of  $p$ . Press “Reveal Bias”. How close was your guess?
5. Press “New Game” to create a new coin. Keep pressing “Flip Coin” until you are confident that you can make a reasonable guess at  $p$ . How many flips did you need? How good was your guess?

**Extra Credit:** Repeat Step 5 multiple time. Do you notice any pattern between the value of  $p$  and the amount of time it takes to make a reasonable guess?

In the sheet “Multiple Flips” you are asked to supply a value for  $p$ , say 50%. Then a “simulation” automatically flips the coin 500 times. The graph shows how the ratio  $(\#Heads / \#Flips)$  changes as the number of flips increases. As we expect, because of the law of large numbers, the values represented on the right side of the graph “converge” to  $p$ .

6. Try different values for  $p$  and generate the graphs. Determine whether values closer to  $p=0.5$  take longer or shorter to converge than values like  $p=0.9$  or  $p=0.1$ .
7. Can you explain why the time to convergence for  $p=0.8$  is similar to that of  $p=0.2$ ?

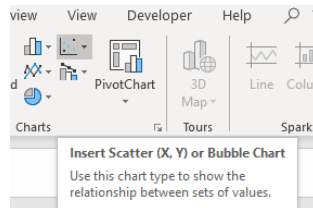
## Relationships between Variables

The file “Corr\_Exercises.xlsx” contains six different paired data sets. For the first data set, where the variables are labeled x1 and y1, we took the liberty of using Excel to construct the “scatterplot” and placing the “trend-line” that best represents any linear relationship that might exist between the variables.

What would indicate that there is a “strong”, “positive”, linear relationship between x1 and y1?

You are going to create a scatter plot for the remaining data sets. Based on the following steps needed to investigate the relationship between x2 and y2

1. In cell B30 type “=correl(A21:A28, B21:B28)”. This provides the value of r, the sample correlations coefficient.
2. Use your mouse to select the range of cells from A20 to B28. Click on the “Scatter Plot” button on the “Insert” Ribbon.



3. Right-click on any of the dots in the plot. Select “Add Trendline....” A dialog box will appear on the right of the screen.
4. For the trendline options, select a “Linear” trendline, select “Display equation ...” and select “Display R-squared ...”
5. Answer the following questions about each of the graphs:
  - a. Which graphs show a strong relationship between x and y?
  - b. Which graphs show a negative relationship between x and y?
  - c. Do any scatterplots indicate that a non-linear relationship between the variables is a better fit? How can you tell?

## Probability Distributions

### Two-way Tables

Partway through the voyage from Southampton, England to New York City, USA the RMS Titanic struck an iceberg and sank in the early morning of 15 April 1912, which resulted in the deaths of about 1,500 people. Data about the survival of passengers and crewmen are summarized in the table below.

	A	B	C	D	E	F
1						
2			Lived	Died	Total	
3		Male	367	1364		
4		Female	344	126		
5		Total				
6						
7						
8			Lived	Died	Total	
9		Male				
10		Female				
11						
12						

This table is in the excel file named “Probability Distributions”.

1. Open the excel file at select the “Titanic Deaths” tab. Use the appropriate spreadsheet formulas to fill in the missing totals in the first table.
2. Place the formula “=C3/E3” in cell C9 to calculate the proportion of men that lived. Use similar formulas to calculate the remaining relative frequencies (percentages) in the second table. Make sure that the values in the table are formatted as percentages.
3. Describe any differences you see between the mortality rate of the Males and the rate of the Females. Do you think a person’s likelihood of surviving was dependent on his or her gender?

The video linked below is an artistic interpretation of the treatment of third-class passengers, but is this attitude towards people in steerage historically accurate? Did the wealthy receive preferential treatment in evacuating the ship?

<https://www.youtube.com/watch?v=Gmw1q0CprEA&feature=youtu.be>

4. Select the tab labeled “Breakdown by Class”. Complete the two-way table on the left of the sheet. Calculate the relative frequencies (percentages) in the table on the right. (Note that the values in every row in column J should sum up to 100%).

5. If one of the passengers is randomly selected from among the first-class passengers, what is the probability that this passenger survived? (That is, what is the probability that the passenger survived, given that this passenger was in first class?)
6. What if the passenger was selected at random from among the third-class passengers?
7. Are your answers to questions 5 and 6 different? What would explain this?

Two-way tables allow us to compare groups and whether they differ from each other in terms of some pair of “variables”, like whether their “Shipping Class” influenced whether or not they “Died” on an unsinkable ship in the North Atlantic during the 1920’s.

The “values” of the variable serve to partition the group based on some category. For example, “Response” can have values {YES, NO}, “Gender” can be broken into {MALE, FEMALE}, or “College Class Standing” can have values {FRESHMAN, SOPHOMORE, JUNIOR, SENIOR}. Other categorical variables can be based on breakdowns like “Color”, or “Demographic Group”, or “Genre”. It is important to make sure that the values of the variables are defined so that each of the subjects that are counted is placed in exactly one category, so 100% of the subjects is accounted for.

### Numerical Data – Single Variable Studies

Sometimes we might want to analyze the breakdown of group in which we have interest by some numerical value rather by category. We might want to see the distribution of values for some variable on the number-line so that we can gain some insight into the population. For example, we may want to analyze the number of children in each family in America, to get a feel for a typical size or to compare these values with those from some other country. We may look at the how much income people make (in \$000 dollars) to get an idea about what constitutes relative poverty, or how much money a plurality of people earns, or how wealth is distributed across the population. In addition, when the variable values are quantities instead of categories, we can apply some mathematical formulas to derive additional information from the data that we are able to collect, (e.g., beyond percentage breakdowns).

8. Select the tab labeled “Mortality Rates, 2017”. The table gives the empirical probability that a person of a certain age currently living in the US will die this year. It is based on data collected during the most recent census. It gives the data for males, females, and for both combined. The expected value for all Americans is 78.08. This means that a person born today has an expected life span of 78.58 years, (where we add .5 to the average because a person doesn’t usually die on his or her birthday but about 6 months after). The sheet also provides the median (or 50<sup>th</sup> percentile age) and the graph of the distribution. What does this information tell you about when a person is likely to die?

9. Notice that the graph has an uptick associated with year 0. Can you provide any possible explanations for this?
10. There is also an uptick associated with year 100. The reason for this is that the category also includes people who died at ages older than 100. What does this tell us about the life expectancy calculation of 78.58?
11. To compare the relative life expectancy of men and women, we can find the expected value of each. In cell H5 place the following formula: “=SUMPRODUCT(B4:B104,C4:C104)”. Use a similar formula to calculate the expected value for females. Does there appear to be a significant difference between men and women, in terms of life expectancy?

**Extra Credit:**

Using the information about the cumulative probabilities, identify the median (50<sup>th</sup> percentile) life span for men and women. What percentage of people in the US are expected to live past 95 years?



## The Normal Distribution

Americans drink an average of 6 cups of water per day. Assume that the number of cups per day for an American is normally distributed with a standard deviation of 1.5 cups.

1. How much water is one standard deviation above the mean?
2. How much water is two standard deviations below the mean?

Suppose we want to find the percentage of the Americans that drink less than 4 cups of water per day.

We will use three different ways to figure out this value.

3. The “Empirical rule” helps us to find an approximate value to the probability that a randomly selected American drinks less than 4 cups per day.
4. The “z- table” allows us to find the probability of interest by hand.
5. The Excel function “NORMSDIST” helps us to calculate probabilities when we have a computer.

Recall from the Empirical rule that: 68% of the data in a Normal distribution is within 1 standard deviation of the mean; 95% is within 2 standard deviations; 99.7% is within 3 standard deviations. Thus the area is distributed as follows.

6. On the x-axis in the picture below, please write in the random variable values associated with the distribution of the amount of water Americans drink per day. (Note that the value 6.0 should be placed in the center.)



7. Place the value 4.0 in the appropriate location in the picture above.
8. Using the picture above, make an estimate to the percentage of Americans that drink less than 4 cups per day.

The z-score of a value of interest represents the number of standard deviations above or below the mean. The formula for the z-score of the value  $x$  is:  $z(x) = (x - \square\square\square\square\square)$

9. Calculate the z-score for the value 4.0 in the distribution of drinking amounts. (Note that this should be a negative number, since 4.0 is less than the mean 6.0).
10. The file “z\_table.xlsx” contains a table detailing the cumulative area of the normal distribution for many values of  $z$ . the value on the left represents the first 2 digits of  $z$ . the columns are associated with the third digit. Find the value of the area associated with the value of  $z$  you calculated in task 4.

The second sheet in the file finds the area using the spreadsheet function called NORM.DIST

	A	B	C	D
1				
2	Mean $\mu$ :			
3	Std. Dev $\sigma$ :			
4				
5	x = key value			
6				
7		=NORM.DIST(B5,B2,B3,TRUE)		
8				
9				

11. Place the value of the mean in cell B2.
12. Place the value of the standard deviation in cell B3
13. Place the value of interest, the key value, (in this case  $x = 4.0$ ), in cell B5.

Your answer will be in cell B7.

Suppose we want to calculate the percentage of the Americans that drink more than 7 cups of water per day. We know that the function gives us the area to the left of the key value. Since we want the area to the right of the value, we subtract the value provided from 1.0.

14. Place the value 7.0 in cell B5.
15. In cell A9 write “Prob( $X > x$ ) =”.
16. In cell B9 write “=1 – B7” to get your answer.

**Exercise:** Use the sheet to find the percentage of Americans that drink between 4 and 7 cups of water per day.

**Consider the following example:**

To qualify for a firefighter position, candidates must score in the top 10% on a general abilities test. The test has a mean of 300 and standard deviation of 30 and it is normally distributed. Find the lowest possible score to qualify.

In this case, we know the desired area, (the upper 10%), but need to find out the key value of the random variable  $x$ .

17. Place the new values of the mean and standard deviation in cells B2 and B3 of the spreadsheet.
18. Place the value 330 into cell B5. Based on the values of the probabilities. Determine whether 330 is too small or too big.
19. Keep adjusting the value in cell B5 until the area below is 90% and the area above is 10% to find the required test score.

Note that you can use the z-table to find the value of  $z$  for which the area to the left is 90%. Find this value.

The third sheet of the excel file provides an easier way to find the desired value.

20. In cell N2 type the value of the mean
21. In cell N3 type the value of the standard deviation
22. Click on the arrows above the Percentile until you reach 90.
23. Look at the values in column Q. What is the required score to pass the test? What is the associated z-score?

## Binomial Probabilities

In this lab, we will explore how to use formulas in Excel to calculate probabilities involving combinations like the “coin-toss” experiments modeled by the Binomial Distribution.

1. How many ways can we select two different letters from the set, {A, B, C, D, E}? List them.
2. What if we selected two from the first 10 letters in the alphabet, {A, ..., J}? Do you think that the number of ways would be much larger? Would it be double or more than double?

It turns out that the number of possibilities practically quadruples when the set of choices doubles. This is an example of the “Combination function”, where COMBIN(10,2) gives the number of ways to get a combination of 2 elements from a set of size 10.

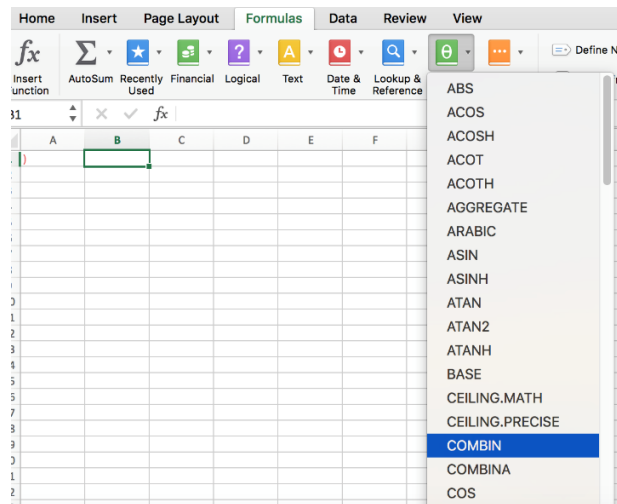
The formula for COMBIN(n,k) is  $n! / k! / (n-k)!$  where  $n!$  is called “n factorial” which is equal to:

$$n! = (n) (n-1) (n-2) (n-3) \dots (3) (2) (1)$$

Use the COMBIN function to find the number of ways to select three different letters from the alphabet:

### Using Excel to Calculate COMBIN(n, k)

- A. Open a new Excel spreadsheet file. Save it as “YourName\_Bin.xlsx.
- B. Move the cursor to cell B1. Click the Formulas icon.
- C. Select the COMBIN function.



- D. Type 10 in the number box.
- E. Type 3 in the number chosen box and hit enter, (or type “=COMBIN(10,3)” into the cell B1)

Use your spreadsheet to answer the following questions:

3. How many different ways can a city health department inspector visit 5 restaurants in a neighborhood with 10 restaurants?

4. How many ID cards can be made if there are 6 digits on a card and no digit can be used more than once?
5. How many different ways can 4 tickets be selected from 50 tickets if each ticket wins a different prize?

Now that we know how to use the COMBIN() function to answer questions on how many ways an event can occur, we can calculate probabilities. Recall that in a “Binomial” experiment, where we want to calculate the probability of getting  $x$  successes from  $n$  repeated independent trials, we use the formula:

$$\text{Prob}(x) = \text{COMBIN}(n,x) p^x (1-p)^{n-x}$$

where  $p$  represents the probability of success in one trial.

(The term “coin flip” model is used when the experiment involves flipping an unfair coin  $n$  times, where the probability of the coin showing heads is  $p$ . We are interested in the probability of showing heads in  $k$  out of the  $n$  flips. Notice that  $p$  is the probability on 1 head when  $n=1$ ).

6. Use Excel to calculate the probability of 3 successes in 10 trials, where  $p = 0.5$

Excel can help us calculate this complicated formula by using the function: BINOMDIST ( )

7. Suppose the president has a 60% approval rating among voters. Twelve voters are randomly selected and asked if they approve of the president. The population is so large that, while sampling is not done with replacement, this does little to affect the independence of the trials. We want to know the probability that seven of the 12 voters approve of the president.

To find the probability follow the steps in your spreadsheet:

- A. In cell A5 type “P =” and in cell B5 type 0.6 (the probability of success).
- B. In cell A6 type “N =” and in cell B6 type 12 (the number of trials).
- C. In cell A8 type “X =” and in cell B8 type 7 (the number we are interested in).
- D. In cell A10 type “Prob(X=7)” and in Cell B10 type “=BINOM.DIST(“
- E. Input the values that the formula requires (“parameters”) as follow:  
Set number\_s to “B8”, set trials to “B6”, set Probability\_s to “B5”, and set Cumulative to 0.
- F. Type “)” and hit enter. The value of the probability will appear in the cell.

	A	B	C	D	E	F
1	P =	0.6				
2	N =	12				
3						
4	X =	7				
5						
6	P(x=7)	=BINOM.DIST()				
7						
8						

What if we wanted to know the probability that at most 5 of the voters approve of the president?

Using the same spreadsheet, we want to calculate  $P(X \leq 5)$

- In cell A11 type “ $P(X \leq 5)$ ” and in Cell B11 type “=BINOM.DIST(“
- Input the values that the formula requires as follow:  
Set number\_s to “B8”, set trials to “B6”, set Probability\_s to “B5”, and set Cumulative to 1.
- Type “)” and hit enter. The value of the cumulative probability will appear in the cell.

In the next example, we will find out how to calculate the probability that at least 8 of the voters will approve of the president. In other words, we want to find  $P(X \geq 8)$ . To find the probability we will need to find the complement probability of  $P(X < 8) = P(X \leq 7)$ .

- In cell A12 type “ $P(X \geq 8)$ ” and in Cell B12 type “=1- BINOM.DIST(“
- Input the values that the formula requires as follows:  
Set number\_s to “B8”, set trials to “B6”, set Probability\_s to “B5”, and set Cumulative to 1.
- Type “)” and hit enter. The value of the probability will appear in the cell.

Save your excel file for future use. Do not close it yet.

### The Binomial Distribution

Open the excel file named, “BinDistrib”. You can see the probability distribution for  $n \leq 25$  and  $0 \leq p \leq 1$ .

- Set p to 0.6 and n to 12. Use the table and the graph to check the answers in your spreadsheet.

## CONFIDENCE INTERVALS

In this lab, we will build and interpret *confidence intervals* for the mean,  $\mu$  of a population. Recall that, even though we don't know the value of the mean, we can obtain an estimate or a range of estimates from a sample. Obviously, the larger the size of the sample, the more confidence we have in our estimate.

### Part I

1. Open the excel file named, "Conf\_Int". The sheet labeled "Population Data" contains thousands of values from some normal distribution where we do not know the value of the mean or standard deviation. We want to make a reasonable estimate to the mean without considering all of that data.
2. On the "Single Sample" sheet, set the sample size to 20 and click on the "Collect Sample" button. Your sample values will appear in the sheet labeled "Sample". The sample mean and sample standard deviation have been calculated for you. Write down these values.
3. Make a reasonable guess about the population mean.
4. Now click the "Collect Sample" button 19 more times and record the means and the standard deviations for each click. Based on this experience, update your estimate to the population mean  $\mu$ .
5. Change the sample size to 500 and repeat Steps 2 - 4. What do you notice about the mean values? What do you notice about the standard deviations? Do you have more confidence in your guess with this larger sample size? Why?
6. Do you think it will be better to give a single number estimate for  $\mu$  or an interval of values? What would give you more confidence about your guess: a larger range or a smaller range?

### Part II

From Part I we should have noticed that sample means are themselves random variables and those coming from a larger sample size have less variability. We will dig deeper by looking at the distribution of sample means.

1. Click on the "Multiple Samples" tab.
2. Set  $n$  to 20 and let the system generate 1000 different samples of size 20.
3. How did the mean of the sample means compare with your previous guess?
4. Do the same with  $n=500$ . What happened to the standard deviation of the sample means?
5. What are the "95% confidence intervals" for  $n=20$  and  $n=500$ ? Can you explain why they are different?
6. The sheet "Histogram" shows the distribution of sample means for  $n=100$ . What do you think the distributions will look like when  $n=25$  instead? What about  $n=500$ ?

## Hypothesis Testing

In this lab, we will test hypotheses about the mean  $\mu$  of a distribution. A null Hypothesis  $H_0$ , e.g., that  $\mu$  is equal to some value  $x$  is believed to be true unless the values from a valid random sample are so far away from  $x$  that we have no choice but to reject the hypothesis and accept some alternative, more reasonable hypothesis.

We will run a simulation that allows us to use all of the steps necessary for the hypothesis testing and drawing a reasonable conclusion based on the sample that we have collected.

The process of any hypothesis test consists of four basic steps:

- Define the null hypotheses  $H_0$ .
- Collect the data: We need random samples that are representative of the population.
- Assess the evidence: Assessment includes checking appropriate conditions, computing test statistics, and finding corresponding P-values. (Where P represents the probability of getting the data we did, given that the null hypothesis is true. The smaller this value is, the less confidence we have in the hypothesis).
- State the conclusion: We compare the P-value to our acceptable level of error  $\alpha$ , and if P is smaller, then we reject  $H_0$ , and state our conclusion.

1. Open the excel file named, "HypTest.xls".
2. Click on the "Test" tab
3. Press the button to "Generate a new population". This gives a population where we do not know the mean or standard deviation. We are asked to test some null hypothesis. Write down this Hypothesis on a sheet of paper.
4. Place a value in cell H3 to represent the size of the sample you want to retrieve from the population. Use  $n=50$ .
5. Click on the "Get Sample" button. The spreadsheet will calculate the sample mean and sample standard deviation. Every time you click the button you will get a different random sample from the same population, so you will get different values for the sample mean and sample standard deviation. (This is why these values are considered random variables).
6. A value for the (t-distribution) test statistic is also calculated for your sample. Write this value on your sheet of paper. Is it positive or negative? If the statistic is positive and the hypothesis is that  $\mu > x$ , or if the statistic is negative and the hypothesis is that  $\mu < x$ , then you can conclude that the Hypothesis is most likely correct. Otherwise go to the next step.
7. Find the associated P-Value for statistic (without its sign) by looking it up in the table located under the tab, "P-Value". Use the column where the number of degrees of freedom is  $(n-1)$ . Write down this value.
8. If the P-Value is less than  $\alpha$  (or  $\alpha/2$  if the hypothesis is  $\mu = x$ ) then you can conclude that the null hypothesis is not likely, so reject it in favor of an alternative hypothesis. Write down your conclusion.