Remarks: Some of the questions on this homework are similar to questions you will find on the midterm exam. Obviously, on the mid-term, you will not be asked to do things in Excel so you can guess which questions these are. Many of the questions ask you to create tables and graphs using MS Excel. For those students unfamiliar with Excel, there are tutorials located on the course homepage demonstrating how to perform these operations.

For those students who prefer to use software other than Excel, that is fine. However, I want you to learn some basic functionality in Excel because you will probably be required to use it in future classes at Chicago Booth.

**Question # 1. Creating tables and histograms**

a. Consider the British marketing data set (bmrbxl.xls) and the “inc” variable, which denotes a household’s income. Make a count table of the “inc” variable. The easiest way to do this is to use the Pivot Table tool in Excel. Hand in a copy of the count table. (Note: a count table for the variable “soc” was shown in the lecture notes for week #1.)

b. Using the same data as in question # 1, create a histogram of the “inc” variable. Hand in a copy of your histogram. (Note: Excel can create many different styles of histograms. You may want to play around with this feature to get comfortable using Excel.)
Question # 2. Creating histograms and time series plots

a. Consider the data on monthly asset returns for different countries (conret.xls). Using Excel, make a histogram of the asset returns for “japan.” You will need to play around with the size of each bin in order to make the histogram look good. Hand in a copy of your histogram.

b. Create a time series plot for ‘japan.’ Hand in a copy of the time series plot.

Question # 3. Creating a time series plot

a. Consider the US real gdp data (realGDPgrowth.xls) available on the course homepage. Create a time series plot. Hand in a copy of the time series plot.
Question # 4. Reading a histogram

Consider three countries from the return data set (conret.xls), which we will label as \( x_1 \), \( x_2 \), and \( x_3 \). Histograms of these series are given below where the horizontal axis is the same for all three:

![Histograms of x1, x2, and x3](image)

The sample statistics for these countries are.

<table>
<thead>
<tr>
<th></th>
<th>( n )</th>
<th>mean</th>
<th>median</th>
<th>std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>germany</td>
<td>107</td>
<td>0.01290</td>
<td>0.01000</td>
<td>0.05612</td>
</tr>
<tr>
<td>usa</td>
<td>107</td>
<td>0.01346</td>
<td>0.02000</td>
<td>0.03328</td>
</tr>
<tr>
<td>finland</td>
<td>107</td>
<td>0.00991</td>
<td>0.00000</td>
<td>0.07800</td>
</tr>
</tbody>
</table>

a. Which of \( x_1 \), \( x_2 \), and \( x_3 \) is the “usa”? Which is “finland”? (NOTE: For this question, you should be able to look at the sample statistics such as the standard deviation and compare them to the histograms to determine the answer.)

b. What is the sample variance of Germany?

c. Using the summary statistics above, give a numeric range that you would expect to contain about 95% of the “usa” returns.
Question # 5. Using Excel to calculate sample statistics

The purpose of this question is to teach you the following functions in Excel. You will likely need them for future classes at Chicago Booth.¹

1. =AVERAGE(...) This function calculates the sample mean of a series.

2. =STDEV.S(...) This function calculates the sample standard deviation of a series.

3. =VAR.S(...) This function calculates the sample variance of a series.

4. =COVARIANCE.S(...) This function calculates the sample covariance between two series.

5. =CORREL(...) This function calculates the sample correlation between two series.

Consider the same three countries from the return data set (conret.xls) as in the last question.

a. Using Excel, calculate the sample mean, sample median, and sample standard deviation for each country. See that they match the results from the table in the previous question.

b. Using Excel, calculate the sample covariance between “usa” and “finland.”

c. Given the results from parts (a.) and (b.), calculate the sample correlation between “usa” and “finland.”

d. Using the CORREL function in Excel, check your results from part (c.).

¹First, notice that in Excel you can always go to the “Formulas” tab at the top. From here, you can click on a specific function to see what it does. If you do not know how the function works, you can click on the help button and Microsoft typically gives an example of how to use it.
Question # 6. Working with summation notation

Suppose we observe the following data

\[ x_1 = 2, \ x_2 = 14, \ x_3 = 8, \quad y_1 = 6, \ y_2 = 2, \ y_3 = 18 \]

a. Calculate \[ \sum_{i=1}^{3} x_i^2. \]

b. Calculate the sample means \( \bar{x} \) and \( \bar{y} \).

c. Calculate \[ \sum_{i=1}^{3} (x_i - \bar{x})(y_i - \bar{y}). \]

Hand in a copy of your work. Show your work by hand (i.e. not in Excel although you can check your answers that way).
Question # 7. Computing summary statistics

a. Consider the bank data (Bank.xls) which measures the interarrival time of customers at a bank. Compute the sample mean, sample median, and sample standard deviation of this data using Excel. Hand in a copy of the output. (Instead of using the individual functions as in Question 5, try using the “Data Analysis” toolbox. Go to the course homepage and find the tutorial on how to do descriptive statistics. It gives a link on how to install this toolbox. The toolbox may be useful in future courses at Chicago Booth and of course at work.)

b. Are the sample mean and sample median of this data very different? Explain why. Plot a histogram to see why. (You do not have to turn the histogram in.)

Question # 8. Using the empirical rule

a. Consider the bank data (Bank.xls) which measures the interarrival time of customers at a bank. Using the sample mean and sample standard deviation of this data, compute a 95% interval for this data. Show how to compute this using the Empirical Rule and hand it in.

b. Next, use Excel’s =PERCENTILE(...) function to calculate the 2.5% interval. How good is the Empirical Rule’s approximation? (NOTE: The bank interarrival data is not mound shaped.)

c. Consider the “usa” variable from the data on country asset returns (conret.xls). In question 4 (c) above you used the empirical rule to compute a 95% interval for this data. Use Excel to compute (approximate) the 95% interval for this data. How good is the approximation? Better than the approximation for the bank data from part (b.) of this question? Why? Hand in your work.
Question # 9. Creating a Two-Way Table

• Pick one of the television shows from the British Marketing survey data (bmbrxl.xls).

• Pick one dummy demographic variable (sex, cola, restE, or cigs) and make a table in Excel that relates viewership of your show to the demographic variable.

• Pick one demographic variable that is ordered categorical or numeric (soc, income, edu, or age) and make a table in Excel that relates viewership of your show to the demographic variable. You should experiment with different groupings (click PivotTable > Group and Show Detail > Group...) to best display the general pattern in the data. For each demographic variable, experiment with different settings (percent of total/row/column, etc.) until you find a table you think best displays the relationship between viewership of the show and that variable (this is of course subjective). Think about how this information might be relevant to marketers, policy makers, network executives, etc. Hand in one of your tables.
Question # 10. Creating a Scatterplot

Consider once again the variable “germany” from the country returns data set (conret.xls).

a. Choose a country other than “germany” whose returns you think should be related to it. Plot the returns against each other in a scatterplot.

b. Choose a country whose returns you think should be unrelated to “germany” and plot these against one another in a scatter plot.

c. Were the countries you guessed correct? Hand in one of the plots of your choice and provide a short explanation why you chose that country.
Question # 11. Reading a Scatterplot, computing covariances, and understanding linear combinations

I have chosen three returns from the monthly mutual fund data (mutualFundReturns.xls), where I have relabeled the returns on each asset as \( x_1 \), \( x_2 \), and \( x_3 \). Some simple summary statistics are

<table>
<thead>
<tr>
<th></th>
<th>( n )</th>
<th>mean</th>
<th>std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>162</td>
<td>0.0035</td>
<td>0.0588</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>162</td>
<td>0.0071</td>
<td>0.0332</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>162</td>
<td>0.0021</td>
<td>0.0184</td>
</tr>
</tbody>
</table>

and the sample correlation matrix is

\[
\begin{pmatrix}
 x_1 & x_2 & x_3 \\
 x_1 & 1   & \text{\hspace{1cm}} & \text{\hspace{1cm}} \\
 x_2 & 0.692 & 1   & \text{\hspace{1cm}} \\
 x_3 & -0.077 & 0.012 & 1   \\
\end{pmatrix}
\]

Figures 1 and 2 are scatter plots of two combinations of \( x_1 \), \( x_2 \), and \( x_3 \) but the graphs intentionally do not indicate which figures are which.
a. Which plot (figure 1 or 2) is a scatter plot of $x_1$ versus $x_2$?

b. What is the sample variance of the returns $x_1$?

c. What is the sample covariance between the returns $x_1$ and the returns $x_2$?

Suppose you had a portfolio which put half your money in asset $x_1$ and half in asset $x_2$. (For parts d. and e. you will need the formulas that we discussed in class.)

d. What would be the average return on your portfolio?

e. What would be the variance of the returns on your portfolio?

f. Give an interval that would contain approximately 95% of the returns on your portfolio.
Question #12. Computing linear combinations and correlation matrices

Consider the monthly return data on portfolios from different countries (conret.xls).

a. Consider a portfolio which invests 10% of wealth into australia, 30% into denmark, and 60% into malaysia. This question asks you to compute the sample mean and sample standard deviation of the portfolio in two ways. In the first method, simply add a column to the spreadsheet that shows the return on the portfolio in each month. (To do this, type the formula $R_p = 0.1r_{australia} + 0.3r_{denmark} + 0.6r_{malaysia}$ into Excel in the top cell in this column. Then, click on the lower right corner of the cell and drag it downward to fill in the rest of the column.) Using Excel formulas, compute the sample mean and sample standard deviation of the portfolio returns. After comparing the sample means and standard deviations of the portfolio to the sample means and stand. dev. of the individual assets, can you explain what you find? Hand in the sample means and sample standard deviations of the portfolio. You don’t need to turn in your Excel output, just write down the answers.

b. In the second method, compute the sample means and sample standard deviations of the three countries’ returns and all the sample covariances between pairs of countries using Excel. Then use the formulas provided in the lecture notes to get your answers. (NOTE: you should get the same thing as in part a.)

c. Make a table of sample correlations for the following six countries: belgium, france, germany, hongkong, malaysia, singapore. Do you see a pattern in this table? (Hint: look at the upper left versus the lower left areas). Hand in the table.
Question # 13. Sample statistics and linear functions

Consider the student data set (studenthw.xls), which has two columns providing the height (measured in inches) and body weight (measured in pounds) of MBA students.

a. What is the sample standard deviation of the variable height?

b. Make a scatterplot of height vs weight. What is their sample correlation? What is their sample covariance? (Again, use Excel or other software and just write down the answers. You do not need to turn in the scatterplot.)

c. Suppose we want the height measured in feet rather than inches. Write down the linear equation that relates height in feet to height in inches. In other words, for the linear equation $H_f = c_0 + c_1 H_i$, where $H_f$ denotes height measured in feet and $H_i$ is height measured in inches, what are the coefficients $c_0$ and $c_1$? Show your work.

d. Find the sample standard deviation of height in feet using your answers to parts a. and c.. Show your work but you may use Excel to check your answer.

e. In Excel, create the column of height values in feet. (NOTE: The easiest way to do this is to click on cell C2 on the spreadsheet, type “=A2/12” (without the quotes), and press enter. Now select cell C2 again, click on the lower right corner of the cell, and drag downward to ‘fill in’ the formula you just entered.) Make a scatterplot of the height in feet values vs the height in inches. What is the correlation between the height in feet and the height in inches? Make a scatterplot of height in feet vs. weight in pounds. What is the correlation between the height in feet and the weight in pounds, and how does it compare with the correlation between height in inches and weight in pounds? (You do not have to turn in the scatterplots.)

f. Find the sample covariance of height in feet and weight in pounds using your answers from parts d. and e. above. Show your work. (You may check your answer in Excel.)

g. Suppose we have two variables $x$ and $y$ with sample correlation $r_{xy}$. And, suppose we define two more variables,

$$ w = a + b \ast x \quad (1) $$

$$ z = c + d \ast y \quad (2) $$
where \( a, b, c, \) and \( d \) are constants. (Note that equations (1) and (2) are linear functions, i.e. the variable \( w \) is a linear function of the variable \( x \). In part c., we saw that height in feet can be written as a linear function of height measured in inches.) What is the sample correlation between \( w \) and \( z \)? I’m not looking for any math here, just use your intuition from the previous parts of this question. Be careful, what happens if \( b \) and/or \( d \) is negative?
Question # 14. Linear regression

This question is about linear regression. Even though we won’t cover regression formally until later in the semester, I want you to get a sense of what we are building toward.

Consider the data on housing prices and size (housesp1.xls). We have observed $n = 128$ houses in the same residential community that were sold in the same year. For each house, we observe the size (in square feet) and the closing price at which the home sold (in dollars). The two columns are the same prices and sizes we saw in Lecture # 2.

a. Run the regression of price on size “by hand”. In other words, use the formulas for the slope and intercept coefficients that we saw in class. To do this, use Excel to calculate the sample mean and sample variance of each variable and the sample covariance between them. Then, plug these values into the formulas for ‘slope’ and ‘intercept’ from the lecture notes. Remember we’re treating price as ‘y’ (the dependent variable) and size as ‘x’ (the explanatory variable or predictor).

b. Now run the same regression using Excel. (There is a tutorial available on the course homepage.) You will need this output to answer more questions below.

c. (This question will not be on a mid-term and you don’t need to turn anything in for it.) In the regression output, there is a number called R-square. Obtain the sample correlation between size and price and square it. Compare that to R-square. That is right, R-square is the sample correlation r squared!! This is true when you have one independent variable. R-square will make sense when we do multiple regression (i.e. fit a line with more than two variables), but it isn’t the sample correlation squared anymore.

d. Suppose you are a real estate agent. Some clients have just called you and hired you to sell 20 houses in this neighborhood. You have to figure out what is a fair price for these houses. Assume you don’t yet know the size of these houses. What is your prediction for the price of the houses? (Hint: Given what you know now, you will have to use the same value as your prediction for each of the twenty houses.)

e. Download the file (housesp2.xls) from the course webpage. It has the sizes of the twenty houses. Now how would you predict the corresponding (but still unknown) house prices? Give a one-sentence explanation and your predictions for the first three houses.