

ETC1000 Business and Economic Statistics
Homework Week 2
(to be completed before your Computer Lab class in Week 3)

This homework is designed to help you summarise numerical data by:

- Constructing a frequency distribution table and cumulative frequency distribution table
- Constructing and interpreting histograms
- Calculating descriptive statistics using Excel formulae and interpreting them

You will also uncover patterns in bivariate categorical data by:

- Calculating joint, marginal and conditional probabilities from a pivot table
- Check for independence in characteristics of interest
- Use Excel to calculate probabilities from the normal and students t distribution
- Compare descriptive statistics across groups

A. Tables and Charts for Numerical Data

Pivot tables and bar charts are usually used for categorical data. They are not generally appropriate for numerical data because numerical data often takes many different values which could be cumbersome, if not impossible, to list. The equivalent to the pivot table and bar chart for summarising numerical data (continuous or discrete) is the frequency distribution table, and its visual display, the histogram. Essentially, the difference between a pivot table and frequency distribution table (and bar chart and histogram) is that the frequency distribution table uses interval ranges, while the pivot table uses categories.

Interval ranges or groupings (bins) may be obvious or not so obvious: for example, the bin ranges for income may be obvious if there are well-defined income groups such as low-income, middle-income, high-income. But if this categorisation is not predefined or is subjective, then it may be up to the analyst to determine what the cut-off value is for each income class.

A frequency distribution table is the tabulation of bin ranges and their corresponding frequency of occurrence in the dataset. The histogram is a graph of this frequency distribution table, with bins on the horizontal axis and frequency on the vertical axis.

This homework will go through an example to illustrate how to use Excel to create a frequency distribution table, cumulative frequency distribution table and histogram.

1. Frequency Distribution Table

The data to be analysed in this homework are results from a class of first year students (found in the *Marks* worksheet in the “Homework 2.xlsx” file). We will use Excel’s *Data Analysis, Histogram* tool to analyse the student results.

First, we need to decide on our bin ranges – that is, what groupings will be displayed on the x-axis of our histogram. For this data, the obvious class (bin) range is that corresponding to High Distinction (HD), Distinction (D), Credit (C), Pass (P) and Fail (N). Enter the following data into your spreadsheet:

	A	B	C	D	E
1	Student ID	Final Mark		Bin Label	Upper Limit of Bin Range
2	1xxxxxx1	65		N	49
3	1xxxxxx2	44		P	59
4	1xxxxxx3	74		C	69
5	1xxxxxx4	60		D	79
6	1xxxxxx5	61		HD	100
7	1xxxxxx6	64			
8	1xxxxxx7	74			

Note that we specify a single value for each bin – the upper limit of the bin range. Here, we have defined the HD bin as a Final Mark of 80-100, a D as 70-79, a C as 60-69, and so on. Note that we can specify an integer value for our upper limit because our data only takes integer values.

Now we use Excel's *Tools, Data Analysis, Histogram* tool. Excel's *Histogram* tool will take our specified bin range and automatically construct our frequency distribution, cumulative frequency distribution and histogram in one go.

On the *Data* tab, select *Data Analysis* from the *Analysis* group (if you do not have the Data Analysis tool installed, see last week's homework for instructions on how to install it).

A pop-up box will appear – select *Histogram* from the list and click *OK*. A second pop-up box will appear. Complete it as follows:

Histogram ? X

Input

Input Range:

Bin Range:

☒ Labels

Output options

☒ Output Range:

☐ New Worksheet Ply:

☐ New Workbook

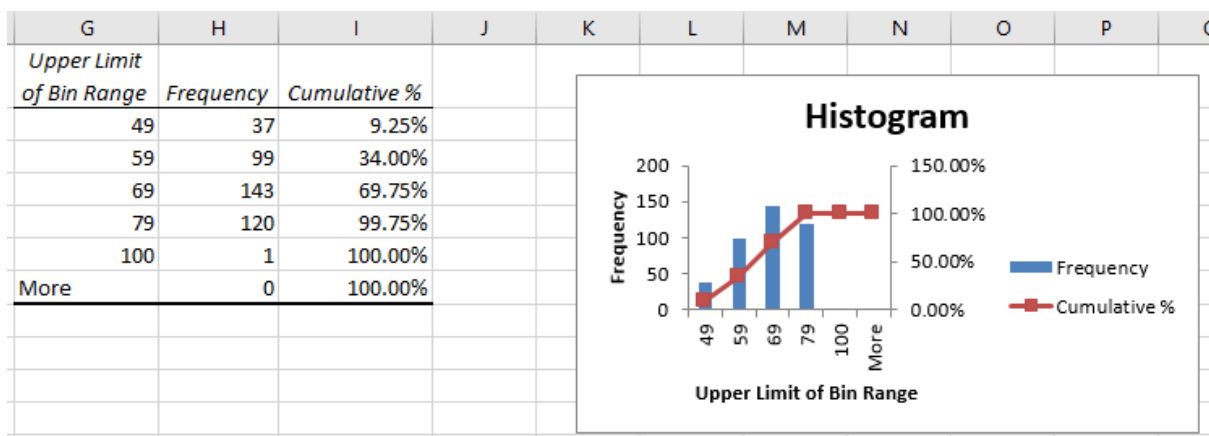
☐ Pareto (sorted histogram)

☒ Cumulative Percentage

☒ Chart Output

- *Input Range*: This is the cell range of your data.
- *Bin Range*: This is the cell range of the bins you specified in your spreadsheet. If you leave this blank, Excel will make up its own bin ranges, but often these are not nice round numbers so are not as clear to the reader.
- We tick the *Labels* checkbox because our input ranges include the names of the data we entered.
- Ticking *Cumulative Percentage* tells Excel to produce a cumulative frequency distribution table as well as a regular frequency distribution table.
- Ticking the *Chart Output* tells Excel to produce a histogram from the frequency distribution.

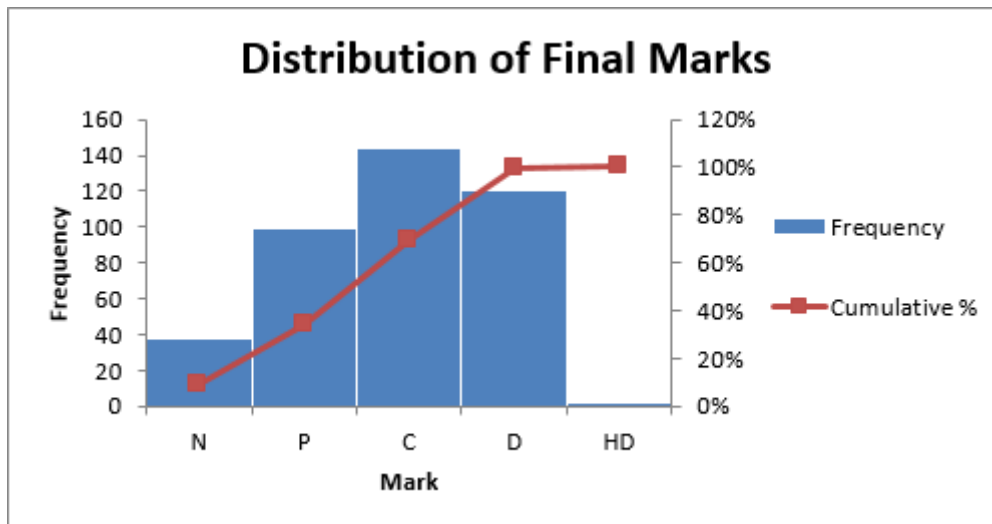
Here is the output obtained when you click OK:



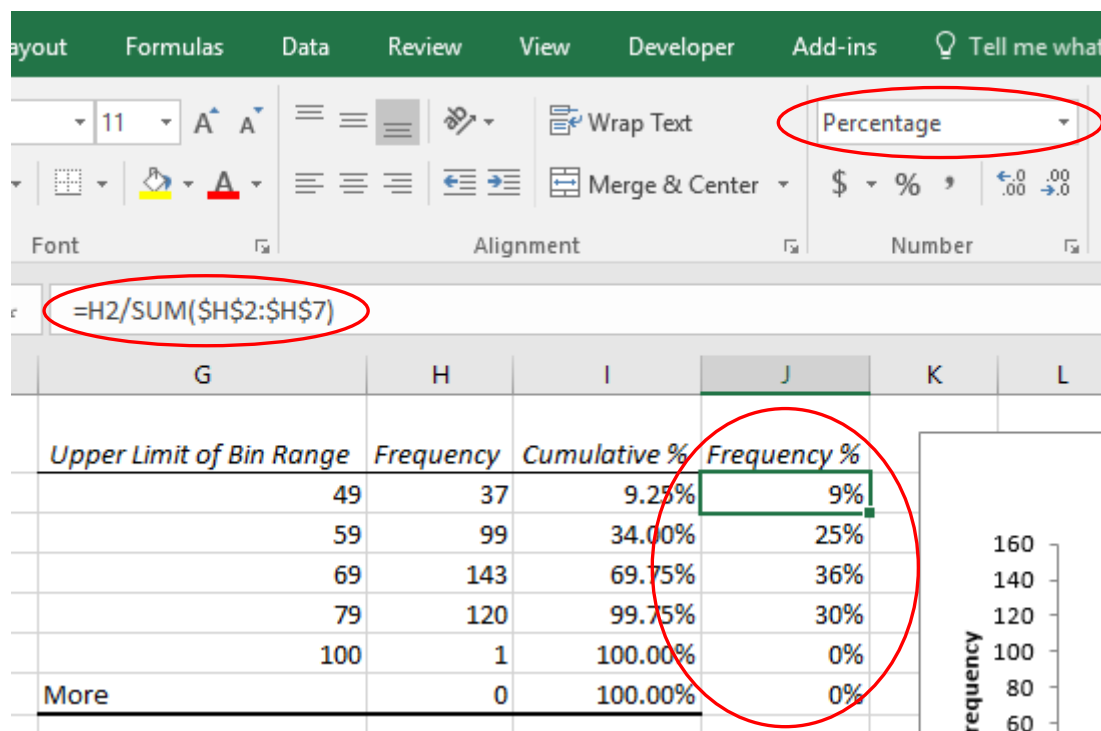
From this output, we can see that 37 of the 400 students failed, and only 1 student achieved an HD – a tough unit!

The presentation of the default histogram is rather poor – it is not helpfully labelled, there are gaps between bars (the class ranges of histograms are continuous by construction, unlike bar charts for categorical data), and the chart area is far too small. But just like a regular Excel chart, we can alter its presentation:

- Enlarge the size of the entire chart area so that the chart plot area is in better (larger) proportion to the legend.
- Remove the “More” category from the chart: right-click on the blue bars for Frequency, and choose *Select Data*. Select the Frequency series and click *Edit*. Change the cell reference to =Marks!\$H\$2:\$H\$6, so that it doesn’t include the “More” bin. Do the same for the Cumulative % series.
- Change the x-axis labels so they reflect the bin labels (HD, D, C, P, N) instead of the upper limit values.
- Change the titles to something more meaningful to the reader.
- Reduce the number of decimal places on the secondary y-axis to make the axis values clearer.
- Remove the gaps between bars: right-click on the blue bars for Frequency, and *select Format Data Series*. Under *Series Options*, reduce the *Gap Width* to 2% (it is visually more appealing to have some gap, even though technically there should not be).

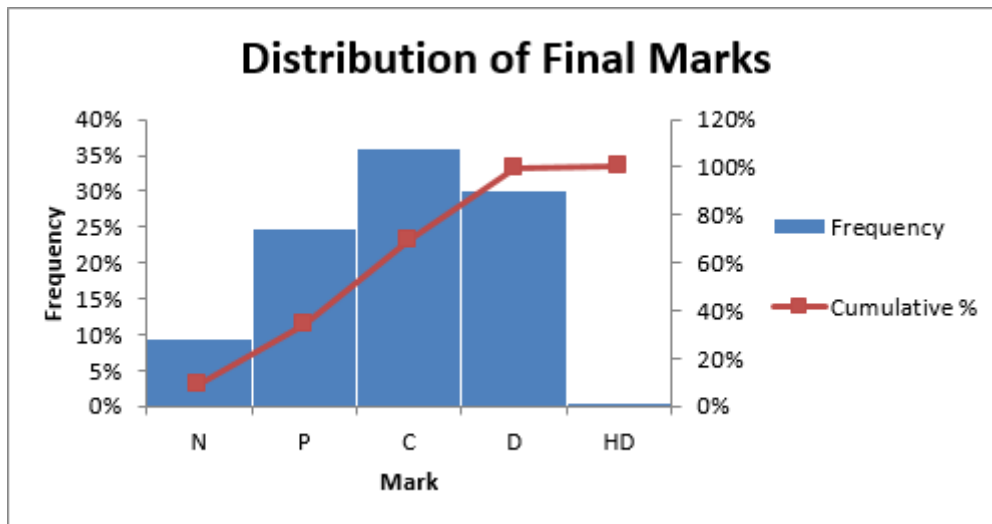


Often it is more meaningful to specify the y-axis in percentage terms rather than number (in this case, percentage of students rather than number of students). To do this we need to convert the frequencies into percentages. Do this as follows:



(N.B. Format cells as Percentage)

Presented as percentage, it is now much clearer to see how the student results are distributed across the grades. And here is the final product:



B. Descriptive Statistics

1. Descriptive Statistics with Excel Formulae

We can calculate descriptive statistics using Excel formulae. Check that you can obtain the following figures for the student results data:

Statistic	Excel function	Result
Mean	=AVERAGE(B2:B401)	62.38
Median	=MEDIAN(B2:B401)	64
Maximum	=MAX(B2:B401)	86
Minimum	=MIN(B2:B401)	13
Range	=MAX(B2:B401)-MIN(B2:B401)	73
Standard Deviation	=STDEV.S(B2:B401)	10.33

Note that the mean is smaller than the median, indicating negative skewness – indeed the histogram in part A shows a long tail of very low results.

2. Descriptive Statistics with Excel's Data Analysis Tool

Excel's Data Analysis add-in will automatically produce a table of summary statistics for you. From the *Data* tab, select *Data Analysis*, and then *Descriptive Statistics* from the list. Complete the dialog box as follows:

Descriptive Statistics ? X

Input

Input Range:

Grouped By: ☒ Columns ☐ Rows

☒ Labels in first row

Output options

☒ Output Range:

☐ New Worksheet Ply:

☐ New Workbook

☒ Summary statistics

☐ Confidence Level for Mean: %

☐ Kth Largest:

☐ Kth Smallest:

Check that you get the same values as those you calculated with Excel's formulae!

C. Extension

Analysis of the student results showed a very low proportion of HDs. Suppose the lecturer in charge decided that they would round up all the students sitting on a mark of 79 up to a mark of 80 – an HD. Let's see what this would do to the distribution of results and the descriptive statistics.

Copy the results data to a new sheet (we do this so we still have the original results if we need to go back to them). In the new sheet, apply a filter to the data to show only the students with results of 79:

	A	B
1	Student I	Final Ma
164	1xxxx163	79
185	1xxxx184	79
326	1xxxx325	79
458		

There are only 3 of them! Now change these 79s to 80s:

	A	B
1	Student I	Final Ma
164	1xxxx163	80
185	1xxxx184	80
326	1xxxx325	80
458		

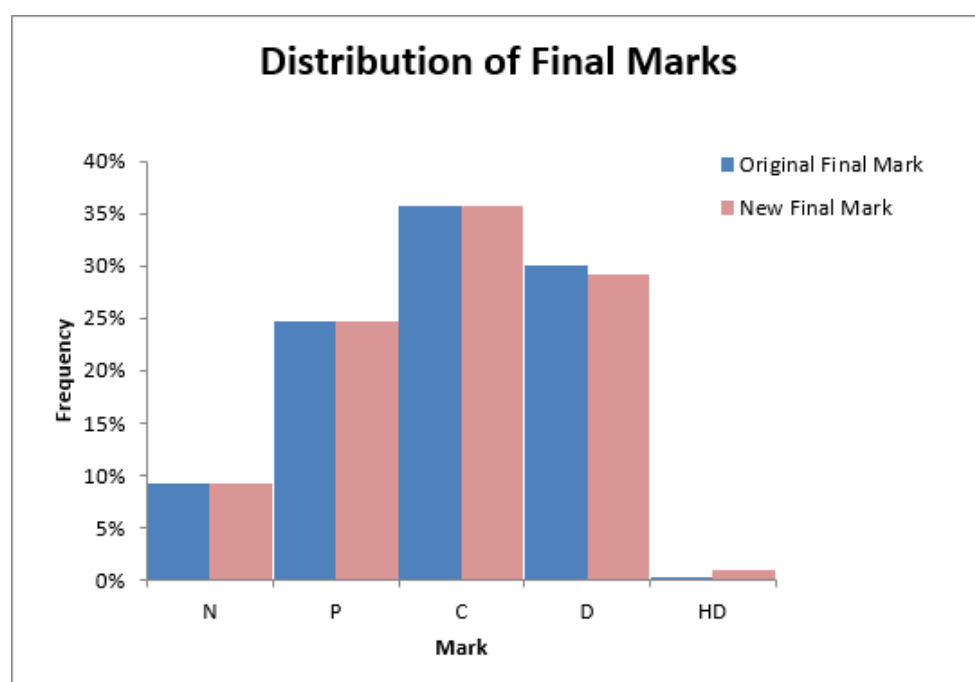
Now re-do the analysis and see how the distribution of the results and the descriptive statistics change. You should find the following:

Statistic	Original Result	New Result
Mean	62.38	62.39
Median	64	64
Maximum	86	86
Minimum	13	13
Range	73	73
Standard Deviation	10.33	10.34

Since there are only 3 students out of the 400 affected by this rounding-up, we only see very slight changes in the mean and standard deviation.

In fact, we should not have expected to see any change in the other statistics, as by definition they would not be affected by the rounding of 79 to 80. Have a think about why this is the case!

Hint: First think about what changing all the 79s to 80s would do to the histogram: a small part of the students in the D category would move into the HD category. How does this affect each of the statistics?



D. Probabilities using a Pivot Table

The Human Development Index (HDI) is an index produced by the United Nations that ranks countries from highest to lowest human development. The index rank is determined based on a number of different measures of development of an economy – based on measures of life expectancy, education and income.

The worksheet labelled *HDI & Schooling* provides the 2013 rankings for countries with complete data on the relevant development measures, geographical region within which the country is located, along with average years of schooling.

1. Probabilities from a Pivot Table

Use Excel's Pivot Table tool to show Region on the rows and the HDI category on the columns as follows:

Count of HDI Category	Column Labels				
Row Labels	High	Low	Medium	Very High	Grand Total
East Asia & Pacific	2	3	15	7	27
Europe & Central Asia	15		5	31	51
Latin America & Caribbean	19	1	10	3	33
Middle East & North Africa	9	2	6	3	20
North America				2	2
South Asia	1	4	3		8
Sub-Saharan Africa	1	36	8	1	46
Grand Total	47	46	47	47	187

Before we look at probabilities using this table, note that the column order is alphabetical, which is confusing, as these categories have a (different) natural order:

Low → Medium → High → Very High

To rearrange the columns so they are more logically ordered (move the *High* column to between *Medium* and *Very High*), click on the black border of cell B4 and drag and drop it into position between *Medium* and *Very High*:

	A	B	C	D	E	F
1						
2						
3	Count of HDI Category	Column Labels				
4	Row Labels	High	Low	Medium	Very High	Grand Total
5	East Asia & Pacific	2	3	15	7	27
6	Europe & Central Asia	15		5	31	51
7	Latin America & Caribbean	19	1	10	3	33
8	Middle East & North Africa	9	2	6	3	20
9	North America				2	2
10	South Asia	1	4	3		8
11	Sub-Saharan Africa	1	36	8	1	46
12	Grand Total	47	46	47	47	187

Now we are ready to begin!

1.1 Table of Joint and Marginal Probabilities

Joint and marginal probabilities can be obtained through a % of Grand Total table. To present the pivot table data in this form, right-clicking on the data in the pivot table and select *Show Data As, % of Grand Total*.

The marginal probabilities are in the 'margins' of the table, highlighted by the solid rectangles below. The joint or intersection probabilities are within the dashed rectangle.

Count of HDI Category	Column Labels				Grand Total
Row Labels	Low	Medium	High	Very High	Grand Total
East Asia & Pacific	1.60%	8.02%	1.07%	3.74%	14.44%
Europe & Central Asia	0.00%	2.67%	8.02%	16.58%	27.27%
Latin America & Caribbean	0.53%	5.35%	10.16%	1.60%	17.65%
Middle East & North Africa	1.07%	3.21%	4.81%	1.60%	10.70%
North America	0.00%	0.00%	0.00%	1.07%	1.07%
South Asia	2.14%	1.60%	0.53%	0.00%	4.28%
Sub-Saharan Africa	19.25%	4.28%	0.53%	0.53%	24.60%
Grand Total	24.60%	25.13%	25.13%	25.13%	100.00%

Joint Probabilities (dashed rectangle around the body of the table)

Marginal Probabilities (solid rectangles around the bottom and right margins)

For example, we can say:

Marginal probabilities:

- $\Pr(\text{East Asia \& Pacific})=0.1444$
- $\Pr(\text{Sub-Saharan Africa})=0.246$
- $\Pr(\text{Very High HDI})=0.2513$

Notice that the column totals are all around 0.25. This is because the United Nations has intentionally placed 25% of countries in each HDI category.

Joint or intersection probabilities:

- $\Pr(\text{East Asia \& Pacific} \cap \text{Medium HDI})=0.0802$
- $\Pr(\text{Sub-Saharan Africa} \cap \text{Low HDI})=0.246$
- $\Pr(\text{Very High HDI} \cap \text{North America})=0.0107$

1.2 Table of Conditional Probabilities

Conditional probabilities can be obtained through a % of Row or a % of Column table, depending on what you are conditioning on. To condition on what is in the rows, use a % of Row table; to condition on what is in the columns, use a % of Column table.

% of Row

First let's condition on what is in the rows – that is, condition on Region. To do this, right-click on the data in the pivot table and select *Show Data As, % of Row*.

Count of HDI Category Row Labels	Column Labels				
	Low	Medium	High	Very High	Grand Total
East Asia & Pacific	11.11%	55.56%	7.41%	25.93%	100.00%
Europe & Central Asia	0.00%	9.80%	29.41%	60.78%	100.00%
Latin America & Caribbean	3.03%	30.30%	57.58%	9.09%	100.00%
Middle East & North Africa	10.00%	30.00%	45.00%	15.00%	100.00%
North America	0.00%	0.00%	0.00%	100.00%	100.00%
South Asia	50.00%	37.50%	12.50%	0.00%	100.00%
Sub-Saharan Africa	78.26%	17.39%	2.17%	2.17%	100.00%
Grand Total	24.60%	25.13%	25.13%	25.13%	100.00%

Here we see probabilities conditional on region:

- $\Pr(\text{Medium HDI} \mid \text{East Asia \& Pacific})=0.5556$
- $\Pr(\text{Low HDI} \mid \text{Sub-Saharan Africa})=0.7826$
- $\Pr(\text{Very High HDI} \mid \text{North America})=1.000$

That is:

- Among East Asia & Pacific countries, the probability a country has a medium HDI is 0.5556.
- Among Sub-Saharan African countries, the probability a country has a low HDI is 0.7826 – very high.
- The probability a North American country has a very high HDI is 1.

% of Column

Now let's condition on what is in the columns – that is, condition on HDI category. Our pivot table using a % of Column display is as follows:

Count of HDI Category	Column Labels				
Row Labels	Low	Medium	High	Very High	Grand Total
East Asia & Pacific	6.52%	31.91%	4.26%	14.89%	14.44%
Europe & Central Asia	0.00%	10.64%	31.91%	65.96%	27.27%
Latin America & Caribbean	2.17%	21.28%	40.43%	6.38%	17.65%
Middle East & North Africa	4.35%	12.77%	19.15%	6.38%	10.70%
North America	0.00%	0.00%	0.00%	4.26%	1.07%
South Asia	8.70%	6.38%	2.13%	0.00%	4.28%
Sub-Saharan Africa	78.26%	17.02%	2.13%	2.13%	24.60%
Grand Total	100.00%	100.00%	100.00%	100.00%	100.00%

Here the probabilities are conditional on HDI category:

- $\Pr(\text{East Asia \& Pacific} \mid \text{Medium HDI})=0.3191$
- $\Pr(\text{Sub-Saharan Africa} \mid \text{Low HDI})=0.7826$
- $\Pr(\text{North America} \mid \text{Very High HDI})=0.0426$

That is:

- Among countries in the medium HDI category, the probability a country is from the East Asia & Pacific region is 0.3191.
- Among countries in the low HDI category, the probability a country is from the Sub-Saharan Africa region is 0.7826.
- Among countries in the very HDI category, the probability a country is from the North American region is 0.0426.

1.3 % of Row vs % of Column?

Notice that conditioning on row and conditioning on column give some very different probabilities. For example:

$$\Pr(\text{Very High HDI} \mid \text{North America})=1.000 \text{ \textit{but} } \Pr(\text{North America} \mid \text{Very High HDI})=0.0426.$$

Conditioning on the rows gives us a very high probability, but conditioning on the columns gives us a very low probability. How do the interpretations differ?

The % of Row tells us that, if we have a list of all the countries with very high HDI, and select one country from the list at random, the probability this country is in the North American region is 0.0426. On the other hand, the % of Column says, if we have a list of the countries from the North American region, the probability a country chosen at random from this list has very high HDI is 1. So while all North American countries have very high HDI, there are only 2 of them (see original table in number display), so they make up a very small proportion of the countries in the very high HDI category (recall that, by definition, approximately 25% of countries are in the very high HDI category).

It can be difficult to work out whether data is best presented as % of Row or as % of Column. In this particular case we would be most interested in % of Row (i.e. $\Pr(\text{Very High HDI} / \text{North America})=1$), as each geographical region has a different number of countries within it. Essentially the % of Row table allows us to make standardised comparisons across regions (the rows). Alternatively, we could think of geographical location as being the 'causal' factor affecting HDI.

2. Convert Frequencies into Percentages by Formula

A pivot table can only be used if there is raw or observation-level data. But often we don't have the raw (e.g. country-level) data to work with, so we need to calculate probabilities ourselves. Suppose, for example, we are researching the HDI performance of Sub-Saharan African countries, and we only have a table in a report and in number display. As mentioned earlier, it would be more informative if the table was displayed as % of Row. How can we convert it? We will need to use formulas in Excel.

First, it will be easier to work with formulas if you copy and paste the data from the pivot table into a new place in the worksheet, without any pivot table links. To achieve this, you will need to copy the pivot table and then right-click and paste values in the new location. Excel will paste the pivot table headings and data in plain text and unlinked:

	A	B	C	D	E	F
1						
2						
3	Count of HDI Category	Column Labels				
4	Row Labels	Low	Medium	High	Very High	Grand Total
5	East Asia & Pacific	3	15	2	7	27
6	Europe & Central Asia		5	15	31	51
7	Latin America & Caribbean	1	10	19	3	33
8	Middle East & North Africa	2	6	9	3	20
9	North America				2	2
10	South Asia	4	3	1		8
11	Sub-Saharan Africa	36	8	1	1	46
12	Grand Total	46	47	47	47	187
13						
14	Count of HDI Category	Column Labels				
15	Row Labels	Low	Medium	High	Very High	Grand Total
16	East Asia & Pacific	3	15	2	7	27
17	Europe & Central Asia		5	15	31	51
18	Latin America & Caribbean	1	10	19	3	33
19	Middle East & North Africa	2	6	9	3	20
20	North America				2	2
21	South Asia	4	3	1		8
22	Sub-Saharan Africa	36	8	1	1	46
23	Grand Total	46	47	47	47	187

We will use this data to construct our % of Row table. First copy and paste the row and column category headings below the duplicate table.

Completing the table with formulas will be much quicker using a mix of absolute cell referencing and relative cell referencing in our formulas before we drag them down or across. Absolute cell referencing is, for example, when you type $\$A\1 rather than just A1 into the formula. Using the “\$” tells Excel not to update the formula reference when it is dragged down or across. You can mix absolute and relative referencing in formulas you wish to drag, such as “=A1/\$B1” or “=A1/B\$1”. To see the result of where the \$ is placed, have a go at dragging down and across the following formulas from cell B27 (you can use the F4 button on your keyboard to toggle these):

- =B16 row and column references change when the formula is dragged
- =\$B\$16 row and column references do not change
- =\$B16 row reference changes, but column reference does not
- =B\$16 column reference changes, but row reference does not

	A	B	C	D	E	F
11	Sub-Saharan Africa	36	8	1	1	46
12	Grand Total	46	47	47	47	187
13						
14	Count of HDI Category	Column Labels				
15	Row Labels	Low	Medium	High	Very High	Grand Total
16	East Asia & Pacific	3	15	2	7	27
17	Europe & Central Asia		5	15	31	51
18	Latin America & Caribbean	1	10	19	3	33
19	Middle East & North Africa	2	6	9	3	20
20	North America				2	2
21	South Asia	4	3	1		8
22	Sub-Saharan Africa	36	8	1	1	46
23	Grand Total	46	47	47	47	187
24						
25		HDI Category				
26	Row Labels	Low	Medium	High	Very High	Grand Total
27	East Asia & Pacific	=B16/\$F16				
28	Europe & Central Asia					
29	Latin America & Caribbean					
30	Middle East & North Africa					
31	North America					
32	South Asia					
33	Sub-Saharan Africa					
34	Grand Total					

Now let's complete the table. For a % of Row table, we want to divide the value in each cell by the row total. That is, $B27=3/27$; $C27=15/27$; $B28=0/51$; $C28=5/51$ and so on. We can calculate these figures automatically using a mix of relative and absolute referencing, starting with the formula = B16/\$F16 in cell B27 as shown above. Now drag this formula across and down to complete the table.

Check that you obtain the same figures that you obtained with Excel's pivot table in % of Row display.

Now try to create the other displays (% of Column and % of Total) using formulae. Hint: if the table above was:

- % of Column, we would use =B16/B\$23.
- % of Total, we would use =B16/\$B\$23.

It is also possible to work backwards from the tables in percentage display to the original number display, and from one table in percentage display to another. Can you work out how?

3. Independence of Events

We can use some probabilities from our tables in percentage display to check whether HDI category is independent of region. For example, if region has nothing to do with HDI category, we would expect that the probability of being in a particular HDI category will be the same across all the regions. For example, conditional on the information of region, the following equality should hold if the two events are independent:

$$\Pr(\text{Low HDI} \mid \text{East Asia \& Pacific}) = \Pr(\text{Low HDI} \mid \text{Europe \& Central Asia}) = \dots = \Pr(\text{Low HDI}).$$

Now:	$\Pr(\text{Low HDI} \mid \text{East Asia \& Pacific})$	=	0.7826
	$\Pr(\text{Low HDI} \mid \text{Europe \& Central Asia})$	=	0.0000
	$\Pr(\text{Low HDI})$	=	0.2460

Since these probabilities are all very different, we would conclude that HDI category is not independent of region.

E. Descriptive Statistics across Groups

How about if the data we wish to compare across regions is numerical, such as expected years of schooling? For example, we might want to compare the amount of education a child in Sub-Saharan Africa is expected to receive compared to a child in the East Asia & Pacific region. Just like we calculated descriptive statistics for a numerical variable of interest, we can calculate descriptive statistics on the numerical variable for sub-groups of the observations – in our case, for each region.

Let's calculate the average Expected Years of Schooling for each region. We can easily do this using a pivot table: first, construct a pivot table placing *Region* in the *Rows* quadrant and *Expected Years of Schooling* in the *Σ Values* quadrant – leave the *Columns* quadrant empty. The pivot table default for this set-up is to count the number of observations (countries) for each region. To change this to display the average for each region, right-click on the pivot table and go to *Summarize values by* and select *Average*. You should have the following:

Row Labels	Average of Expected Years of Schooling
East Asia & Pacific	8.237037037
Europe & Central Asia	10.54117647
Latin America & Caribbean	8.063636364
Middle East & North Africa	6.95
North America	12.8
South Asia	4.725
Sub-Saharan Africa	4.560869565
Grand Total	7.69144385

We could have calculated averages by group using Excel formulae. To do this, copy the region names to a space in the worksheet, and enter the formula in the first cell (G2) as follows:

B	C	D	E	F	G	H	I	J
Region	HDI Category	Expected Years of Schooling			Average			
South Asia	Low	3.1		East Asia & Pacific	=AVERAGEIF(\$B\$2:\$B\$188,F2,\$D\$2:\$D\$188)			
Europe & Central Asia	High	10.4		Europe & Central Asia	AVERAGEIF(range, criteria, [average_range])			
Middle East & North Africa	High	7.6		Latin America & Caribbean	8.063636			
Europe & Central Asia	Very High	10.4		Middle East & North Africa	6.95			
Sub-Saharan Africa	Low	4.7		North America	12.8			
Latin America & Caribbean	High	8.9		South Asia	4.725			
Latin America & Caribbean	Very High	9.3		Sub-Saharan Africa	4.56087			
Europe & Central Asia	High	10.8		Grand Total	7.691444			
East Asia & Pacific	Very High	12						

Notice how we have used the mix of absolute and relative addressing in this formula, so that we can drag down the formula rather than type it in for each region. The formula for *Grand Total*, of course, is simply =AVERAGE(D\$2:D\$188).

N.B. We could, in fact, have referenced the entire columns B and D, that is, use AVERAGEIF(B:B,F2,D:D). This makes working out the formulas easier, and also means if more data is added below row 188, the formulas do not need to be altered.

Try out some of the other “....IF” functions in Excel!

F. Using Excel to Calculate Probabilities from the Normal Distribution

We can use Excel to calculate probabilities from continuous distributions like the normal distribution. To demonstrate, let's consider an example.

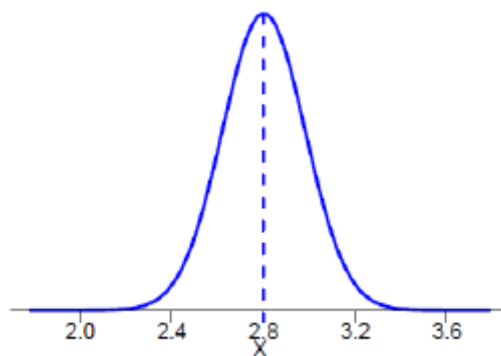
A consumer group surveyed food prices at stores in South Australia. Among the food prices being measured was that of a 2kg bag of sugar. Arguably, the price of sugar could be considered continuous. Suppose that the price of sugar has a normal distribution with a mean of \$2.8 and standard deviation of \$0.18. Use the Excel NORM.DIST or NORM.INV function to answer the following questions:

- (1) What is the probability that the price of a bag of sugar (2kg) from a randomly selected store in South Australia costs more than \$2.80 and up to \$2.85?
- (2) What is the probability that the price of a bag of sugar from a randomly selected store in South Australia costs more than \$2.80?
- (3) Brett has just bought a bag of sugar from a shop in South Australia. What is the probability that he paid more than \$3.00?
- (4) Prices of sugar at stores in one particular suburb in South Australia are said to be in the highest 1.5% of South Australia's sugar price. What is the price range of sugar at stores in this suburb?

1. Excel's NORM.DIST Function

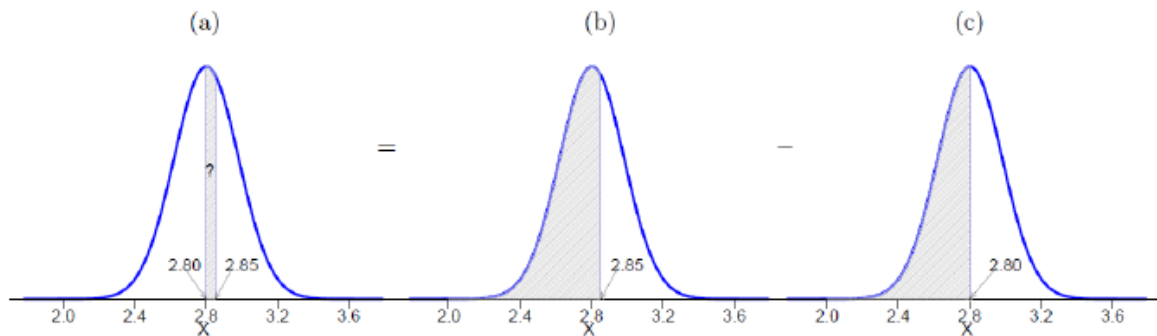
As the price of sugar is considered continuous, there are an infinite number of possible values that the price could take, and the probability of any one particular value occurring is zero. Therefore, we talk about the chances of the price falling within a particular range. In technical terms, we call the sugar price a *random variable* – a quantity of interest that can take a range of possible values.

The way to examine a random variable is through its probability distribution. A probability distribution is the tabulation of events (price ranges) along with their associated probabilities of occurrence. As assumed in the exercise preamble, the price is normally distributed, with a mean of \$2.80 and standard deviation of \$0.18. Denoting the sugar price as X , the graphical representation of the corresponding probability distribution is as follows:



The plot highlights the symmetric shape of the normal distribution: its skewness is zero. The sugar prices cluster towards the middle of the distribution and tail off toward the extremes (positive and negative). The mean and median of a normally distributed variable are the same. This graphical representation of the probability distribution is called the PROBABILITY DENSITY FUNCTION (PDF). By definition, the total area under a PDF is 1.

Question (1) asks the probability that the value of X falls within the range (2.80, 2.85), which is: $\Pr(2.80 < X < 2.85)$. This probability is equal to the shaded area in panel (a) below, which is the area difference between that of (b) and (c). Accordingly, $\Pr(2.80 < X < 2.85) = \Pr(X < 2.85) - \Pr(X < 2.80)$.



How do we use Excel's NORM.DIST function to calculate these probabilities?

For a given normally distributed random variable, the NORM.DIST function gives the probability that the random variable falls below a specified value. The format of this function is *NORM.DIST(x, mean, standard_deviation, TRUE)* with:

- *x*: the specified value to be examined
- *mean*: the mean of the normal distribution
- *standard_deviation*: the standard deviation of the normal distribution

To answer question (1):

Step 1: The probability that the sugar price is less than \$2.85 is:
 $\Pr(X < 2.85) = \text{NORM.DIST}(2.85, 2.80, 0.18, \text{TRUE}) = 0.6094$

Step 2: The probability that the sugar price is less than \$2.80 is:
 $\Pr(X < 2.80) = \text{NORM.DIST}(2.80, 2.80, 0.18, \text{TRUE}) = 0.5000$

N.B. In this exercise, the median of the normal distribution is \$2.80, indicating half the sugar prices are less than this value, therefore the probability of a sugar price less than \$2.80 is 50% (or 0.5). This probability is exactly equal to the answer given by *NORM.DIST(2.80, 2.80, 0.18, TRUE)*.

Step 3: Therefore:
 $\Pr(2.80 < X < 2.85) = \Pr(X < 2.85) - \Pr(X < 2.80) = 0.6094 - 0.5000 = 0.1094$

Now use Excel to answer questions (2) and (3) yourself. You should obtain the following answers:

- (2) 0.5000
- (3) 0.1333

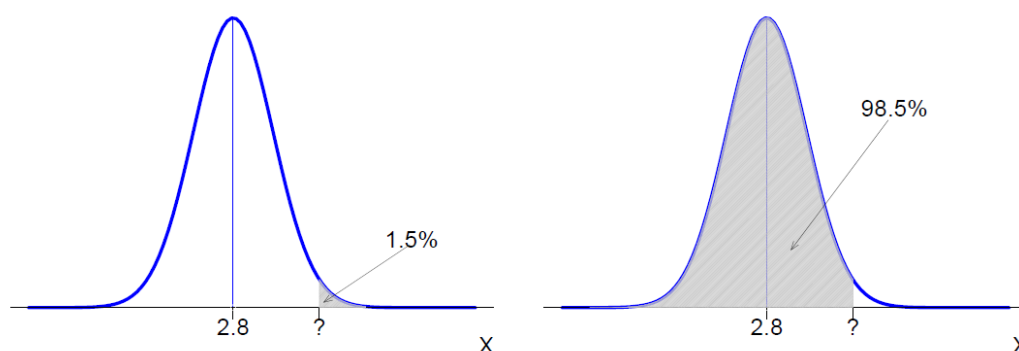
2. Excel's NORM.INV Function

To answer the last question in our exercise, we need to use Excel's NORM.INV function, which is the inverse function of NORM.DIST. NORM.INV calculates, for a given normal distribution, the value of the random variable given a probability, and its format is *NORM.INV(probability, mean, standard_deviation)* with:

- *probability*: the probability to be examined
- *mean*: the mean of the normal distribution
- *standard_deviation*: the standard deviation of the normal distribution

Excel's NORM.INV function assumes the probability we input is the probability in the direction of the left-hand tail, i.e. it calculates “?” in the probability statement “ $\Pr(X < ?) = \text{probability}$ ”. To illustrate, question (4) asks: how much would the price of a bag of sugar need to be if its price is among the highest 1.5% of South Australia's sugar prices? In other words, $\Pr(X > ?) = 0.015$. Note that this is a “greater than” probability. If we were to input NORM.INV(0.015, 2.80, 0.18) then Excel would be calculating $\Pr(X < ?) = 0.015$, which is not what we want. So how can we get Excel to give us the value of X for a “>” probability?

Note that: $\Pr(X > ?) = 0.015 = 1 - \Pr(X < ?)$, which can be re-arranged into: $\Pr(X < ?) = 1 - 0.015 = 0.985$, as shown in the figures below:

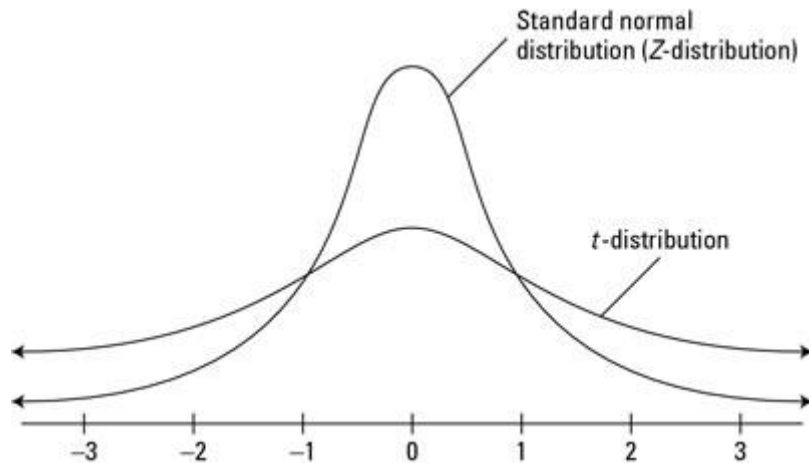


So we can use NORM.INV as follows: $\text{=NORM.INV}(0.985, 2.80, 0.18) = 3.1906$. Therefore, the sugar price needs to be at least \$3.19 for it to be among the highest 1.5% of South Australia's sugar prices.

Extension: Can you use the result of $\text{NORM.INV}(0.015, 2.80, 0.18) = 2.4094$ to obtain the correct answer to this question?

3. Students t Distribution

The Students t distribution (often just referred to as the t distribution) is a continuous distribution that we will use extensively in this unit. The t distribution is centred around zero and is similar in shape to the standard normal distribution (which has mean = 0, standard deviation = 1), but with more area in the extremes (the positive and negative “tails” of the PDF). That is, the t distribution is bell-shaped, but a bit more spread out.



In Excel we use the functions:

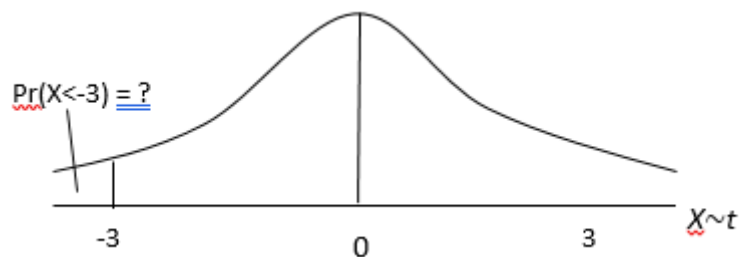
`=T.DIST(X, degrees of freedom, TRUE)` : to work out probabilities for a given value of X.

`=T.INV(probability, degrees of freedom)` : to work out the X value for a given probability.

There are some things to note about T.DIST and T.INV:

- T.DIST and T.INV (and their extensions) require a “degrees of freedom” input, which is its version of a standard deviation: we will explain more about what this actually is later in the unit, for now, just treat it as a given integer value.
- Since the t distribution is symmetrically centred around zero, $\Pr(X < -t) = \Pr(X > t)$.

Let’s use an example to illustrate: suppose $X \sim t$ (degrees of freedom=30), and we wish to calculate $\Pr(X < -3)$.



We use `=T.DIST(-3, 30, TRUE)` = 0.0027. Note that to calculate $\Pr(X > 3)$, we could use the exact same formula, or the `T.DIST.RT(X, degrees of freedom)` function, where “RT” refers to the right-tail of the distribution. Confirm this by calculating `T.DIST.RT(3, 30)` in Excel. The function `T.DIST.2T(X, degrees of freedom)` calculates the probability in both tails of the distribution. That is, if we typed `T.DIST.2T(3,30)` into Excel, it would calculate $\Pr(X < -3) + \Pr(X > 3) = 2 * \Pr(X < -3) = 0.0054$. The reason why T.DIST is set up this way will become clearer later in this unit.

T.INV works in the reverse, and assumes the probability is in the left tail of the distribution. Have a go at calculating `T.INV(0.0027,30)` in Excel to see what you get. T.INV.2T gives a value for t assuming that the probability is split equally across the two tails of the distribution: `T.INV.2T(0.0054, 30) = 3` tells you that $t = \pm 3$ are the values of X for which $\Pr(X < -t) + \Pr(X > t) = 0.0054$. Can you work out what the trick might be to obtain a value of X for a probability in just the right tail?

So how do these values of X and probabilities compare to what we would have gotten from a standard normal distribution? Use Excel to check - calculate $\Pr(X < -3)$, where $X \sim N(\text{mean}=0, \text{standard deviation}=1)$. Compare this to the value you obtained using $T.DIST(-3, 30, \text{TRUE})$. Reconcile how this accords with the difference in shape of the standard normal and Student's t PDFs.

How about using $T.INV.2T$? Calculate the value of X , where $X \sim N(0,1)$, for which 2% of the data lies in the lower tail and 2% lies in the upper tail. Compare this with the value of X you obtain if $X \sim t(\text{d.o.f.}=30)$.