

## Module 3: Quantitative Economics 1- Basic Statistics

---

### Learning Goal for Module 3

After you have defined the problem and acquired/processed the relevant data, the next step in any economic analytics exercise involves data descriptions to begin understanding the relationships among the variables. This Module begins the Level 2 analysis described at the start of Module 2.

By the end of this Module, you will:

- Understand what the terms *variable* and *parameter* mean
- Align the level of measurement, the data available, and the nature of the problem
- Understand the difference among association, correlation, and causation
- Understand the term *economic model*
- Install and use the Data Analysis option
- Create a picture of your data in the form of a distribution
- Describe your data using measures of central tendency and variation.

## 1. Introduction

This Module begins the process of using economic analytics to derive meaning from numbers. Referring to the data analytics ladder of Module 2, the common starting point for any economic research project is preparing a description of the data. Tables and graphs, then summary statistics, begin the journey from association and correlation toward inferring cause and effect, predicting the future course of variables of interest and manipulating causal variables to simulate policy impacts.

For a business, important causal variables include price, expenditures on advertising, or wage rates. For macroeconomic policy, tax rates, social safety nets, monetary policy manipulation of interest rates, and the purchase/sale of assets form variables of interest. For microeconomic policies, prices, costs, and psychological factors are among the more common variables of interest.

Policy variables occupy a special place in economic/business analytics. Changing tax rates or adjusting product prices are supposed to produce specific outcomes implied by our theory and models. The most important goal of economic analytics is to understand the relationships among economic variables to construct practical models that support predictions of what will happen as we adjust the policy variable in advance of making these changes.

**Example:** A pizza business reduces the price of its large pepperoni and cheese pie by 10% because its analysis shows the elasticity of demand exceeds  $-1.25$  and a price reduction could increase total revenue.

**Example:** Government imposes the tax rate on “excess” profits in the bank sector. It expects/hopes that the bank will not pass this tax on to consumers through increased service fees

and that bank management will reduce dividends to shareholders to teach them not to be greedy.

In both cases, the predictions depend on economic models that start by collecting and then describing relevant data as the first step in developing a causal model.

### 3.1. Variables and parameters

Economic models derive from theories that use assumptions, conjectures about human behaviour, and logic to form rules about how the economy and society work.

**Example:** The quantity demanded of any good (or service) is *inversely* related to the price and *directly* related to the income of the household.

Y and X are *inversely* related if when X increase Y decreases Y and X are *directly* related if when X increases, Y also increases

**Example:** “The fundamental psychological law...is that men [and women] are disposed, as a rule and on average, to increase their consumption as their income increases, but not as much as the increase in their income.” (J.M. Keynes, *The General Theory of Employment, Interest and Money*)

These general theoretical statements need translation to statements that use concrete symbols to denote the component variables. The following models use variables for the above two examples:

Example:  $Q_d = f(P, Y)$ , where  $Q_d$  is quantity demanded,  
 $P$  is price, and  $Y$  is income.

Example:  $C = f(Y)$ ,  $\frac{dC}{dY} > 1$ ,  $\frac{d^2C}{dY^2} < 1$

(Consumption increases with income, but at a decreasing rate, which is what Keynes said above.)

A variable is a letter that stands for an attribute, characteristic, number, or quantity that varies (increases, decreases, or stays constant) over time. It may assume different values in different contexts. Used abstractly, in theory, such as in the equations above, a variable uses letters, initials, or acronyms, such as the variables “X” and “Y.” In a concrete model, where the data show sales at different prices of a product, a variable measures a specific quantity sold or the price.


*Example:* Inequality has been falling. (abstract)

*Example:* The Gini coefficient is lower now than 10 years ago. (concrete)

A primary task in economic analytics is to translate an abstract hypothesis or statement into a concrete statement with clearly defined measures.

Variables have four dimensions, based on form and role.

**Levels of measurement** is one attribute of a variable. Table 1 shows the levels of measurement, which is easier than trying to define the idea.

Levels of measurement			
	Ratio data	Fixed distance between measures and a true zero exists	Weight, age, income
	Interval data	Fixed distance between measures, no actual or true zero	Temperature in Fahrenheit
	Ordinal data	Ordered variable	Student letter grades or professor rating (good, average, bad)
	Nominal data	Categories only – no orders	Sex (M or F), homeowner or renter

The orders from higher to lower seems counterintuitive but think of dropping to earth using a parachute. The ability to see detail increases at lower altitudes. Very generally, lower levels of measurement will support improved analysis but will involve increased complication in definition and data collection.

**Causal role** is an attribute of a variable based on the role of a variable in theory or statistical modelling.

**Independent variables** can take different values and are presumed to *cause* corresponding change in dependent variables.

**Dependent variables** can take different values only in response to changes in one or more independent variables.

*Example:*  $C = a_0 + a_1Y$  is the familiar Keynesian consumption function, where changes in Y (income) *cause* changes in consumption spending (C). The dependent variable here is C and the independent variable is Y.

Module 3: Quantitative Economics 1- Basic Statistics

**Example:**  $Q_{di} = a_0 + a_1P_i + a_2P_j + a_3INT$  is an equation for the quantity demanded for product i. The equation shows that the quantity demanded for good i depends on changes in its own price, the price of a second product j, and the interest rate. Another way to say this is that changes in the price of good i, the price of good j, and the interest rate cause changes in the quantity demanded.

Microeconomics theorizes quantity demand to be a function of the price of i ( $P_i$ ), the price of another product ( $P_j$ ) that can be a substitute or complement, and  $INT$  is the interest rate on consumer loans. Here the independent variables are  $P_i$ ,  $P_j$ , and  $INT$ , and the dependent variable is  $Q_d$ .

Special letters  $a_0$ ,  $a_1$ , etc. denote statistical coefficients taking different values depending on the data. Theory will often dictate feasible values for these coefficients. In the consumption function,  $C = a_0 + a_1Y$ ,  $a_1$  will be greater than 0 and less than 1, while in the demand equation, we usually assume that  $a_1$  is less than 0.

**Continuous and discrete variables** are a further category of the variables. Nominal and ordinal data will usually be **discrete** variables and only assume fixed values, while interval and ratio variables are **continuous**, since data can take infinite numbers of values between two points.

Discrete			Continuous		
Type	Example	Values	Type	Example	Values
Nominal	Sex	Male, Female	Interval	Rainfall 20 mm	Cannot be negative.
	Gender	Male, Female, +			
Ordinal	Letter grades	A+, A, B+, B...	Ratio	Income -\$15,230 (standing for a business loss)	Can be negative and high, with unlimited decimal places.
	Scale on survey question	Poor, fair, average, good, very good			

A variable that never changes in value is a **constant**.

A variable that assumes only set values, sometimes defined by the analyst as a **parameter**, is estimated statistically or used in simulations.

### 3.2. Economic and business models

Economic and business models have three formats:

#### *Verbal/text models*

We usually formulate theory using words such as, “Sales are a function of the price of the product (own price), the price of the competitors’ products (cross prices), the income of the consumer, interest rates, etc.” Another example might be, “Using Justin Bieber to sell donuts to young people makes sense because people want to emulate cultural heroes and will buy Biebs Brew and Biebs Bits.”

The box below shows how Keynes formulated the marginal propensity to consume, which underlies the consumption function, although he never wrote out the model mathematically. It was Alvin Hansen, an American economist, who developed this concept shortly after Keynes published the *General Theory of Employment, Interest and Money* in 1936.

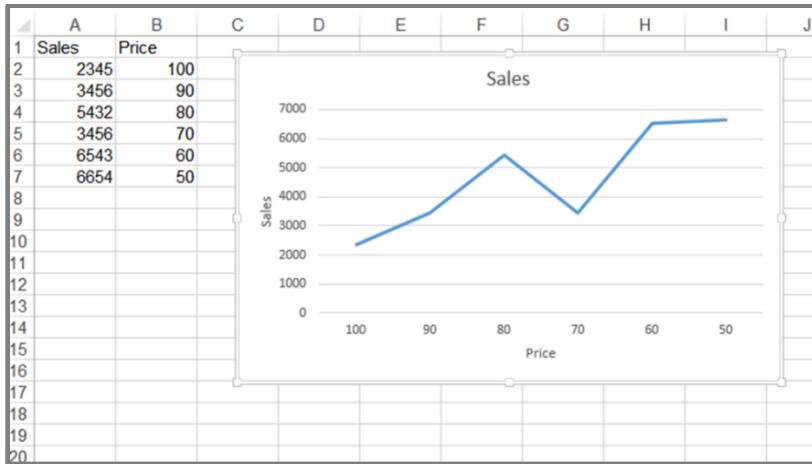
Think of the consumption function defined earlier. Here is a verbal or text version of the model.

“Our normal psychological law that, when the real income of the community increases or decreases, its consumption will increase or decrease but not so fast, can, therefore, be translated — not, indeed, with absolute accuracy but subject to qualifications which are obvious and can easily be stated in a formally complete fashion — into the propositions that  $\Delta C_w$  and  $\Delta Y_w$  have the same sign, but  $\Delta Y_w > \Delta C_w$ , where  $C_w$  is the consumption in terms of wage-units.”

John Maynard Keynes *The General Theory of Employment, Interest and Money*

#### *Graphical and tabular models*

The simple chart of sales and price in Figure 1 below is a graphical model. Graphical models allow one to see data patterns immediately and support first conjectures about theory, but beyond two independent variables, visualization becomes challenging.



This is the demand relation with the axes reversed.

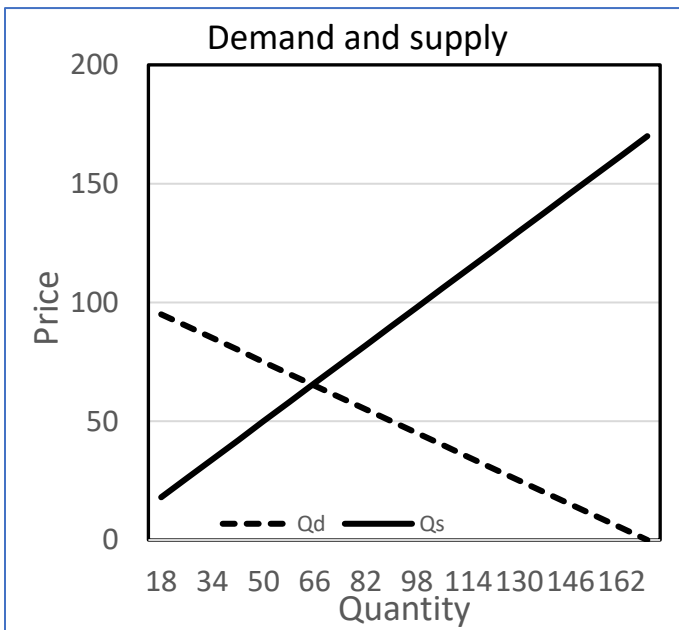
Here sales are the dependent variable (effect), and price is the independent variable (cause).

Figure 1: Sales vs price

**Mathematical models**

The demand and supply diagram is a basic graphical model used in applied microeconomics and easily constructed in Excel. Most people understand that the model includes the demand (Qd) and supply (Qs) equations but omits the equality condition which completes the model

[Demand\\_Supply.xlsx](#)



$Q_d = a_0 + a_1 P$  ....1  
 $Q_s = b_0 + b_1 P$  ....2  
 $Q_d = Q_s$  ....3  
 $a_0, a_1, b_0, b_1$  are all parameters.  $Q_d, Q_s$  and  $P$  are variables (continuous). Equations 1 and 2 are behavioural equations since they express how consumers (1) and suppliers (2) change quantity demanded/supplied in response to changes in price. Equation 3 is an identity and essential to the model.

Figure 2: Demand and Supply

## Module 3: Quantitative Economics 1- Basic Statistics

---

Examples of other models in this course include probability models; statistical models; models of taxation; financial models, including cost-benefit analysis; optimization models, including the linear programming model; and simulation of processes, such as disease progression (COVID). Excel can support applications in all these areas and more.

### *Decoding the language of cause and effect*

A model in economics serves two purposes. First, it is a descriptive summary of a relationship among a set of variables. Second, it can be a causal statement. Now, it is hard to prove cause and effect. Randomized control trials (with participants) are a common way to study cause and effect. Increasingly, economists infer causal relationship from observational or administrative data use quasi-experimental techniques.

The language used by an analyst shows whether they intend the model to stand for a cause-and-effect relationship. The following sentences imply a causal relation – watch out for the key terms that signal a causal idea.

- Quantity demanded *depends* on price.
- Quantity demanded *is a function* of price.
- Price *determines* quantity demanded.
- *If* we increase our advertising budget, *then* we will increase sales (if... then).

### 3. Formulas: Order of calculation and use of parentheses (a deeper dive)

This section comes from the online help function in Microsoft 365 (F1).

#### Arithmetic

Operator	Interpretation	Example	Result
+ (plus)	addition (inverse of subtraction)	=2+4	6
- (minus)	subtraction (inverse of addition)	=3-1 =8-9	2 -1
* (shift 8)	multiplication (inverse of division)	=2*2	4
/ (forward slash)	division (inverse of multiplication)	=6/2	3
% (shift 5)	percent	=20%*4	80% (0.8)
^ (shift 6)	exponentiation	=2^2	4

No extra spaces

Module 3: Quantitative Economics 1- Basic Statistics

Comparison		Reference		
Operator	Example	Operator	Meaning	Example
= (equal)	A1=B1	: (colon)	Denotes all cells between the two named cells	A1:A10 (row) A1:D20 (matrix)
> (greater than)	A1>B1	, (comma)	Combines multiple reference	=SUMPRODUCT(A1:C10,D1:F10)
< (less than)	A1<B1			
>= (greater than or equal to)	A1>=B1			
<= (less than or equal to)	A1<=B1			
<> (not equal to)	A1<>B1			

Order of operations

Operation		Comparison
: (colon)	Reference	=
, (comma)		< >
- (negation)	e.g., -1 or -=A1	<=
%		>=
^	Exponentiation raising to a power	<>
*, /		
+, -		

Order of operator precedence

Formula operations have the following precedence:

1. Evaluate terms in parentheses
2. Perform negation (-)
3. Evaluate ranges (A1:A100, A1:GH1, or A1:F20)
4. Evaluate intersections (spaces)
5. Evaluate unions (,)
6. Convert percentages (%)
7. Perform exponentiation (^)
8. Perform multiplication (\*)
9. Perform addition (+) and subtraction (-)
10. Evaluate text operations (&)
11. Perform comparisons (>,<,>=,<=,<>)

Operations run left to right

Example:  $10^2 = 100$  and  $-10^2 = 100$   
(negation precedes exponentiation)

Good practice

1. Use parentheses extensively to make sure your calculations are correct.  

$$=(A1*B1)+(C1*D1)*E1$$
 is different from  

$$=A1*B1+C1*D1*E1$$
2. Use a calculator to manually check a calculation, especially if you are going to copy the formula across multiple rows.

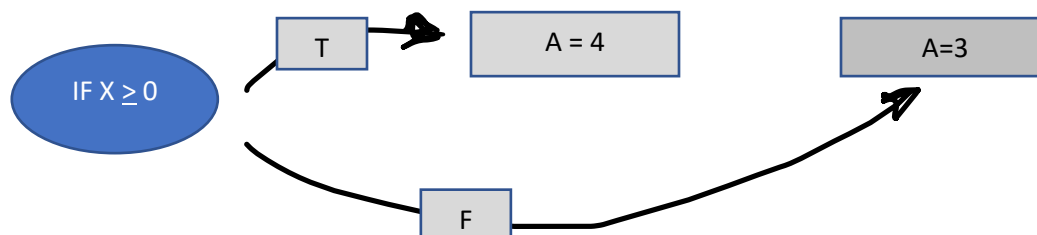


Formula errors can be difficult to understand. Here are seven pointers.

- **#DIV/0!** – dividing a numeric by another numeric with the value of zero
- **#N/A** – occurs when a logical function cannot perform a match or when an argument is missing from a formula
- **#NAME?** – a name or value not recognized
- **#NULL!** – two ranges that do not intersect (overlap)
- **#NUM!** – a problem exists with the formula
- **#REF!** – invalid cell reference
- **#VALUE** – you are pointing to a cell with the wrong type of entry

#### 4. A first look at logical formulas: the =IF formula

The =IF formula is a sentence that state IF “X is true,” then do “Y,” otherwise do “Z.” All computer codes have this branching operation.

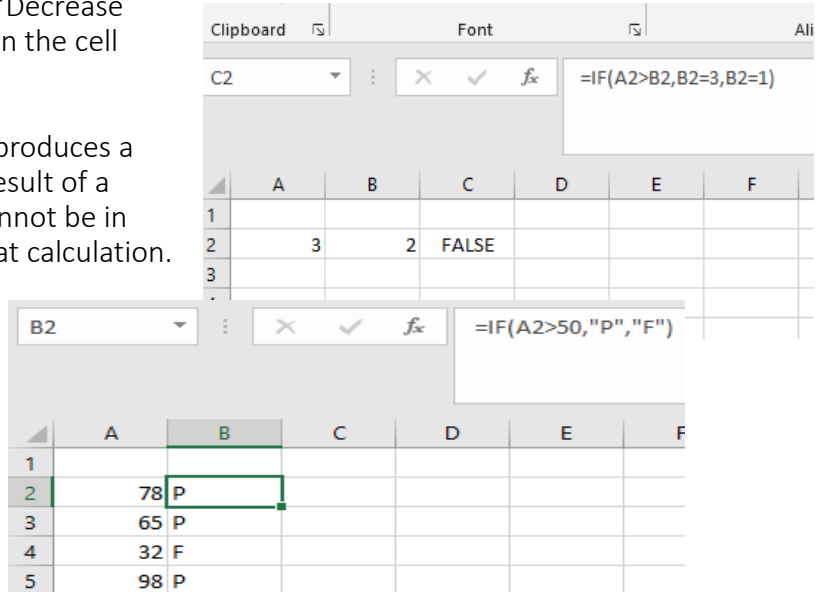


We will use this formula throughout the course. Examples:

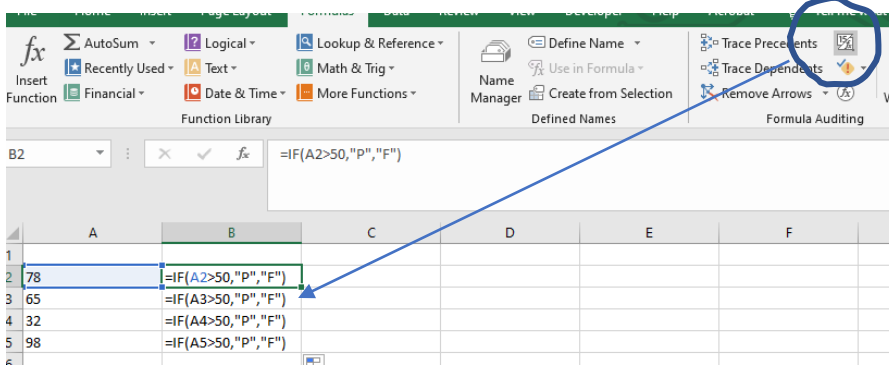
- =IF(A2=1,B2=100,B2=0) – If the contents of cell A2 = 1, then set the contents of cell B2=100, otherwise, set the contents of B2 = 0. We assume B2 stores the results of the IF statement. To ensure “100” and “0” appear as the desired results and not “true or false,” enter only 100 and 0 without the “B2=”.
- =IF(A2=“Fred”, B2=100, B2=0) – IF statements work with text.

**Economic Analytics Using Computer Methods: Econ 2050**  
**Module 3: Quantitative Economics 1- Basic Statistics**

- =IF(A2>B2,“Increase price”, “Decrease price”) –The result appears in the cell where the =IF occurs.
- =IF(A2>B2,B2=3,B2=1) This produces a circular error, because the result of a mathematical calculation cannot be in any of the cells that form that calculation.
- Grade calculation



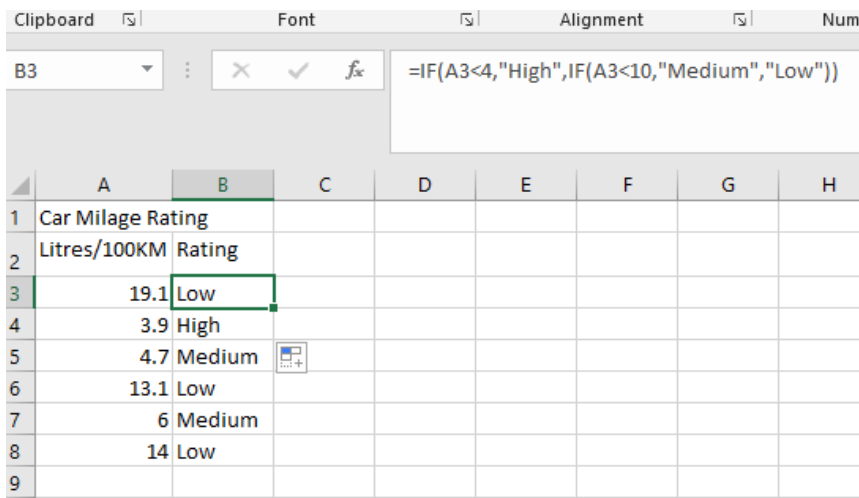
- Show formulas



The “show formulas” option can be useful in debugging formulas.

[NestedIF.xlsx](#)

- “Nested IFs” are especially powerful and support multiple branching. See



Study the logic of the nested IF.

## 5. Frequency distribution

A frequency distribution, usually in the form of a histogram, is often the first “picture” we create of data. We can create histograms manually by first sorting the data and then grouping the entries into “bins.” We can vary the bin size; more bins imply smaller ranges for each bin, and fewer bins means wider ranges.

Common first steps in data analysis include sorting and grouping data. If we have a set of 10 numbers (23, 32, 19, 22, 18, 21, 31, 14, 12, 29), we can sort the list into ascending or descending order, using the Sort function in Excel. We can group the data into ranges, which Excel terms “bins.” Obviously, a longer list with a wider range of data will support more bins. Also, the bin range affects the number of bins supportable by any set of numbers.

Original	Sorted ↑	Sorted ↓
23	12	32
32	14	31
19	18	29
22	19	23
18	21	22
21	22	21
31	23	19
14	29	18
12	31	14
29	32	12

Video: [SORT](#)

Three bins of <19, 20–25, >26 produce the following table. Note how the use of  $\geq$  and  $\leq$  is so important to precisely specifying the bin ranges. In the sort on the left, we have omitted the value “19.”

Bin	n	Bin	n
<19	3	$\leq 19$	4
20–25	3	20–25	3
>26	3	$\geq 26$	3

### 5.1. Histograms

A histogram is the most basic picture available to describe data, grouping information into ascending categories to show a basic *distribution* for the data. When we sort data and create bins, we create a histogram. The most important function of a histogram is to show the frequency distribution of the data. See.

COUNTIF.xlsx

Video: [COUNTIF](#)

=COUNT and =COUNTIF allow you to count cells within a range that meet a specific condition:

=COUNT (counts the number of cells in a range that have numbers)

=COUNTIF (counts the number of cells in a range that meet a condition)

The form of COUNTIF is

=COUNTIF(range, criteria)

=COUNT(A1:A10) counts the number of cells in the range A1–A10 that have numbers. This is useful when you have lists with blanks, weird characters, and letters as well as numbers.

=COUNTIFS extends the COUNTIF to multiple ranges (Look up under F1).

=COUNTIF(A1:A10, "Canada") counts the number of cells with "Canada" in cell A1 through cell A10.

Video: [Histograms using COUNTIF](#)

The Data Analysis ToolPak offers another and sometimes more convenient approach to creating a histogram. You still need to manually create the number of bins and their sizes. See

[Histogram.xlsx](#)

Video: [Histogram with Data Analysis ToolPak](#)

Note: In Chapter 13, the =FREQUENCY command will simplify the creation of histograms.

## 5.2. Histogram to frequency distribution

Joining the midpoints of the histogram bars traces a frequency distribution. This is the basis for thinking about the *data generating process* underlying the data. Sometimes the data generating process might be a natural process, creating probability distributions such as a normal distribution (intelligence), log-normal (the size of natural gas deposits), or Poisson (people arriving at an online store). Chapter 5 looks at these in more detail.

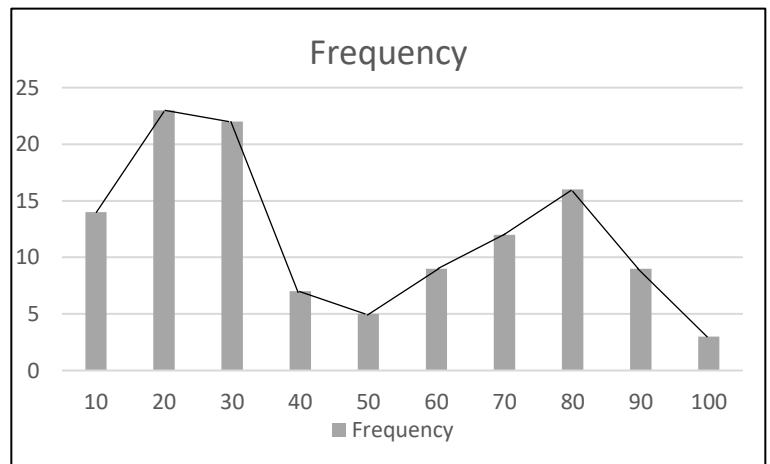


Figure 3: Frequency Distribution

Other times, the frequency distribution may not be associated with an established model or is just the result of random events. Studying the distribution of your data is an important part of Level 1 analysis.

### 5.3. Central tendency and variation

The average is the most common way we summarize economic and social phenomena. The grade point average or GPA summarizes a university career, while the average winter temperature for Winnipeg in January (-19C) summarizes weather severity. The average is one measure of central tendency.

Variation is a basic feature of the natural and social order which reveals unusual people, places, and things. But before we speak of variation, we need a standard against which to measure the common or usual. In any human gathering, we naturally notice the usual and unusual, whether it be height, weight, or other visible differences.

#### 5.3.1. Measures of central tendency

The **mean** or **average** seems a natural way to reference the usual. Some dimensions of difference may be visible, while others are innate and invisible, assessed only through specific measures such as genetic make-up or assessments of intellectual capacity.

Summation notation:

$$\sum_{i=1}^n X_i = X_1 + X_2 + X_3 + \dots X_n$$

The mean is part of a group of **measures of central tendency** comprising the mean, median, and mode, among others.

- The **mean** or “average” (also known as the **arithmetic mean**) represents the centre of gravity for a set of numbers. Imagine you present the data in question as a histogram cut out of plywood, then the mean is the balance point.

$$\bar{X} = \sum_{i=1}^n x_i / n$$

The **weighted mean** offers a convenient way to create mean when certain numbers repeat, or you wish to change the impact of a specific set of numbers on the global mean.

Module 3: Quantitative Economics 1- Basic Statistics

Survey research uses weighted means when samples over- or under-represent an important aspect of the population.

$$\bar{X} = \sum_{i=1}^n w_i x_i$$

The **mode** is the most common value; the **median** is the value that divides the area under the histogram or frequency distribution in half. For the Mincome data, and the distribution of ages of female heads, we have the following results.

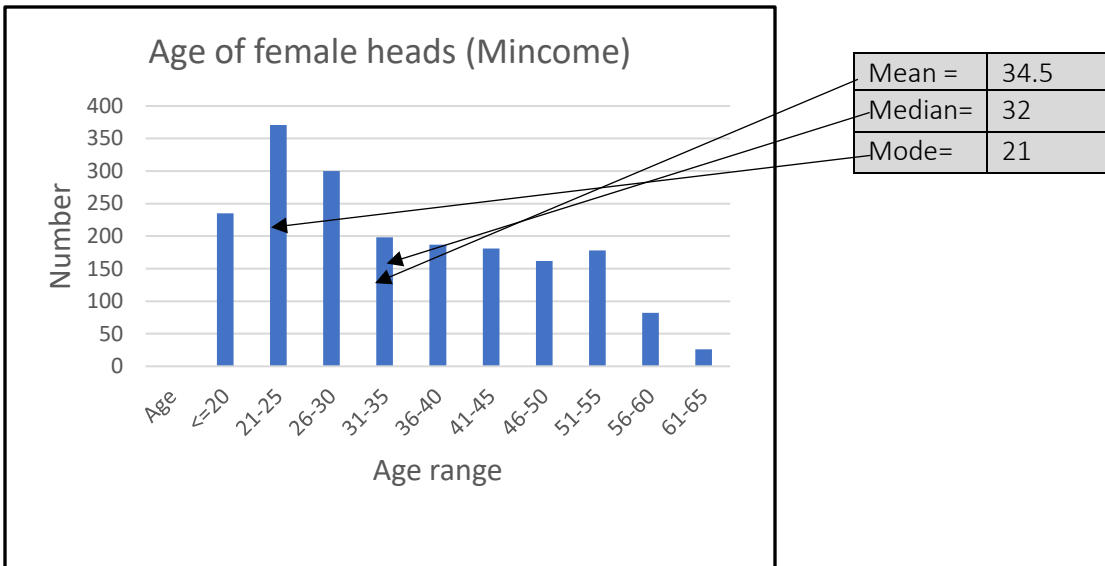


Figure 4: Mean. Median. Mode

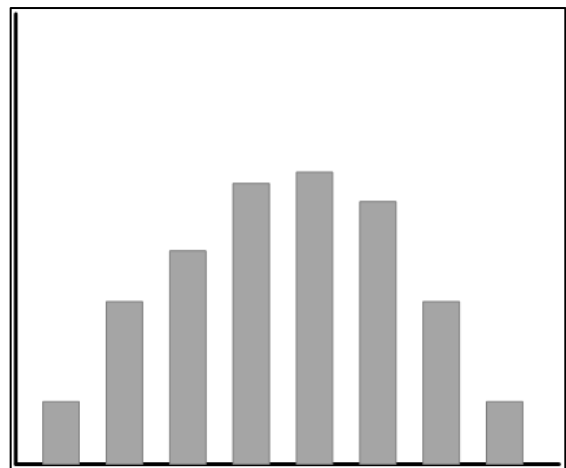
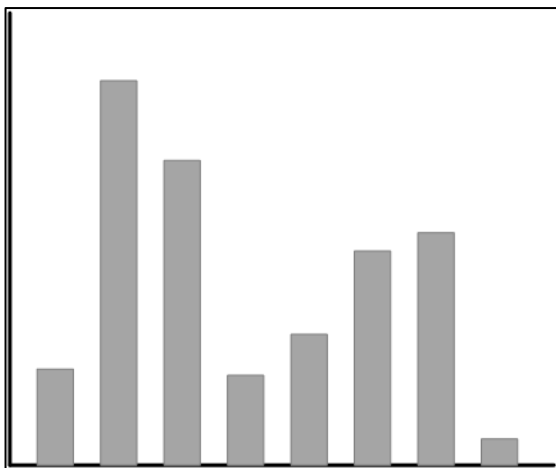


Figure 5: Bimodal and Unimodal?

For a unimodal distribution, the order of measures of central tendency is mean-median-mode for a skew-left distribution and mode-median-mean for a skew-right distribution.

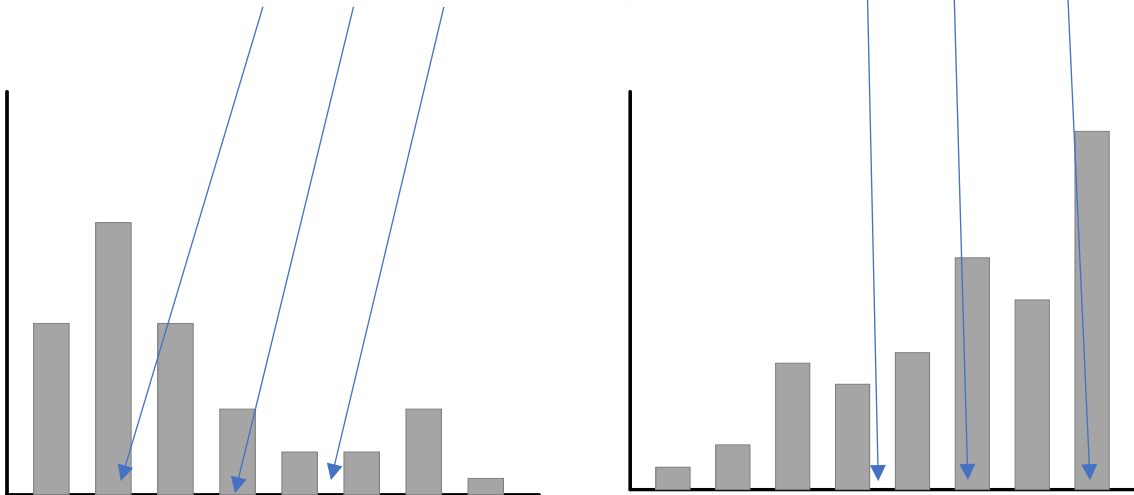


Figure 6 - Skew right and skew left

A histogram categorizes data into fixed steps called *bins*. The ability to infer any details of shape and relative sparseness of the data depends on the number and size of the bins in the histogram. Increasing the number of bins means reducing their range and vice versa. Specifying the range of each bin requires care to avoid having data falling between the bin ranges. Analysts must judge which combination of bin number and bin ranges best suits the goals of analytical insight and communicating information.

The grades in mathematically oriented courses often have a bimodal distribution, where many students do well, but a large group struggle. Symmetric distributions, such as the normal distribution, are common in the natural order and form the foundation of classical statistics. The normal distribution usually characterizes the distribution of human attributes (intelligence).

These are the guidelines for using the mean and median to assess the skewness of distribution of data:

- If the mean and median are close, this usually indicates a symmetrical distribution.
- If the median and mean diverge, this can mean:
  - skewed distribution

- extreme valued observations or outliers, typically due to data entry errors or unusual values that require investigation (*Hint: Never ignore or suppress [delete] outliers without first trying to understand their source.*)
- If the distribution is skewed and data are continuous, use the median as the measure of central tendency (i.e., the typical value) rather than the mean, since outliers (extreme value data) affect means more than medians.

**Example:** For this set of 10 numbers (3, 5, 2, 7, 2, 5, 7, 1, 0, 65) the median is 4 and the mean is 9.7. Assuming 65 is a valid observation, the median of 4 is more representative of the numbers than the mean.

One last point is that the mean or average is the expected value for a distribution of numbers. Imagine an experiment where the GPA of 10,000 randomly selected students appeared on a card in an urn. Let us imagine that you know the average GPA for the school (this is the population value). If you could win \$10 if you guess the value on a card within  $\pm 0.5$ , then using the mean for every guess is your best strategy. You will likely be wrong on any single bet, but after 100 guesses, using the mean would net you the highest level of winnings. That is because the mean is the most likely number from this experiment.

Of course, the dispersion of the GPA scores also matters and this example reappears below.

### **Geometric mean**

Implicit in the arithmetic mean is the assumption that the data points are on a linear scale. For example, using the female age data from Mincome, each year of age is the same “distance” apart. But consider this example. You own a home, purchased initially for \$500,000. It appreciates in value in the first year by 25%, then for the next three years at 10%. In the fifth year, it increases 5%. What is the average annual change in value over the five years? The arithmetic average or mean is  $(25\%+10\%+10\%+10\%+5\%)/5 = 60\%/5 = 12\%$ .



Let us compute actual changes in value. We use Year 0 (January 1) as the date of purchase, and Year 1 as the first complete year of appreciation. The house value grows to \$873,469 by Year 5. The arithmetic mean of the growth rate in each year is 12% per year. Now if we multiply \$500,000 by  $(1.12)^5$  we get \$881,171, which overstates the actual growth. Dealing with percentages, ratios, and exponential growth, the appropriate mean is the geometric mean, defined as:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Year	Rate of change	Year to Year change	Value
0			\$500,000
1	25%	1.25	\$625,000
2	10%	1.10	\$687,500
3	10%	1.10	\$756,250
4	10%	1.10	\$831,875
5	5%	1.05	\$873,469
Average change	12%	1.12	

In the example, we need to take the 5<sup>th</sup> root of  $(1.25 \cdot 1.1 \cdot 1.1 \cdot 1.1 \cdot 1.05)$ , which is 1.118. If we multiply  $(1.118)^5$  times \$500,000, we get \$873,331 which is close; roundoff error creates the imprecision. The geometric mean results in slightly lower value than the arithmetic mean.

This example can be difficult to follow, so please see

[Geometric.xlsx](#)

for a detailed explanation with some extensions.

Cell comments offer documentation and appear as triangles in the upper right of the cell. Position yourself (“mouse over”) in the cell to read. In your own work, right click and choose “Insert Comment.” You can edit/delete a comment by right clicking on a cell and choosing the appropriate option. It is a useful habit to add cell comments to remind yourself of what you did and to communicate with team members.



### 5.3.2. Measures of variation (data dispersion)

The variance is the most common measure of **variation**, which shows the “spread” of the values in our dataset. The wider the spread, the greater the number of extreme observations. Our awareness of “the unusual” depends on a *deviation* from the norm, the commonplace, or most often the mean or median.

A very tall person typically stands out because they deviate from the mean or average. Measuring the deviation of each member of the group from the mean lies at the heart of classical statistics. Explaining variation in data is the core of economic analytics; if no variation exists within a dataset, it is uninteresting.

## **Module 3: Quantitative Economics 1- Basic Statistics**

---

Variation also underlies our ideas of cause and effect since matching the variation in a “cause” with the variation of an “effect” represents an important (but not the only) clue in creating insight into causality. Philosophers debate the concept of causality and the nature of the evidence needed to create the basis for concluding that X is a cause of Y. Intuitively, we expect causes and effects to vary, either:

- directly (an increase/decrease in the value of a causal variable is associated with an increase/decrease in the value of the effect variable), or
- indirectly (an increase/decrease in the value of a causal variable is associated with a decrease/increase in the value of the effect variable).

Another important idea of causality is that causes occur before effects. An event may have multiple causes, and we say that X is a necessary but not sufficient cause for Y when other events need to occur to trigger the occurrence of Y. Suffice to say that, in this course, we offer only a superficial treatment of causality.

Typical causal questions include:

- Will drinking coffee after 3 p.m. allow me to get a better night’s sleep?
- If I lose 10 kilos, will I be more successful in online dating sites?
- Will reducing our advertising in newspapers and increasing the use of social media increase sales? Should I use Facebook, Instagram, or TikTok to increase sales among 15- to 20-year-old women?
- Will lowering the bank rate increase investment and lead to increased GDP and incomes? How will negative interest rates affect savings?
- Will increasing the tax on gasoline reduce driving and/or cause consumers to switch to more fuel-efficient cars and/or encourage consumers to increase their use of transit, and will this reduce greenhouse gas emissions?

The variation is the spread in a frequency distribution. The most basic measure is the range or the difference between the lowest and highest values, which, for the female headed households in the Mincome data, is 65–15, or 50. The range is a crude measure, but it is always useful to record the lowest (=MIN) and largest values (=MAX) just to make sure you have no unusual data points. Computing the difference between the maximum and minimum values will produce the range.

Recall that Mincome used -9 as an indicator of missing data, and so had you used the function =MIN on the original data, it would have shown instances of -9, which is a strange value for age. Obtaining a true idea of the mean, median, and variance of female ages requires that we trim (or clean) the missing data from the file.

The *variance for a population* uses the following formula:

$$\sigma_x^2 = \sum_{i=1}^N (X_i - \bar{X}) / N,$$

where N is the population size and  $\bar{X}$  is the population mean.

The *variance for a sample* includes an adjustment for losing a *degree of freedom* and appears as:

$$\sigma_x^2 = \sum_{i=1}^n (x_i - \bar{x}) / (n-1)$$

where n is the sample size and  $\bar{x}$  the sample mean. (Note that uppercase X denotes the population and x denotes the sample, while uppercase N denotes the population size and n denotes the sample size.)

The *standard deviation* is the most common representation of the variance and is simply the square root of the variance. It too has a population and a sample version. When applied to the distribution of parameters in a statistical model, the standard deviation is often the *standard error*. See [Variance.xlsx](#)

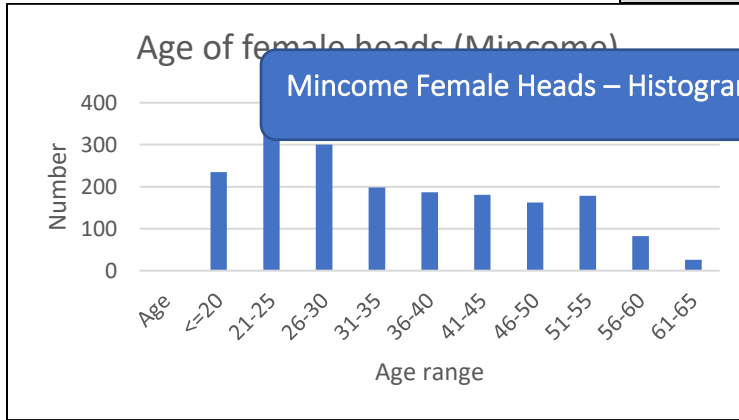
The concept of *degrees of freedom* is an important and difficult idea in statistics, a detailed discussion of which is beyond this course. A simple presentation uses the variance definition from the equation above. The essence of the idea is that any set of numbers has a story to tell. Describing that set of numbers consumes information. For example, if I have the set {1, 2, 3}, one part of the story is the mean, which equals 2. Now if I gave you the set {X, 2, 3} and told you the mean of the set was 2, there is only one choice for X, namely 1. The action of computing the mean uses one degree of freedom.

So why divide the sample variance by n-1 and the population variance by N? This has to do with the concept of bias in sampling,

Here is a simple calculation of the variance and standard deviation of four numbers

i	$x_i$	$x_i - \bar{X}$	$(x_i - \bar{X})^2$
1	4	1	1
2	5	2	4
3	1	-2	4
4	2	-1	1
Sum	12	0	10
		$\sigma^2$	10/3
		$\sigma$	$(10/3)^{1/2}$

Variance	153.2259
Standard deviation	12.37844



Mincome Female Heads – Histogram, Mean, Variance.xlsx

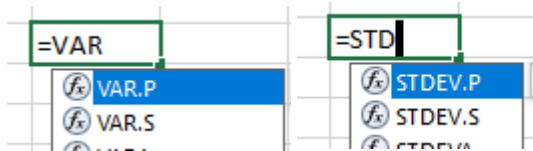
Because the variance is often large and standard deviation relates to symmetric distributions, analysts prefer its use as a measure of variation.

See

Mincome Female Heads- Histogram, Mean, Variance.xlsx] and [Mincome Female Heads- Percentile].

Figure 7: Age distribution of female heads - Mincome

Excel offers options for variance and standard deviation.



As discussed above, VAR.P and STDEV.P calculate the value for a population, while VAR.S and STDEV.S apply to a sample. *When in doubt, stick to VAR.P and STDEV.P.*

The variation in time series data often reveals trends and seasonal variations that are useful for creating forecasting models, as shown in Figures 4 and 5.

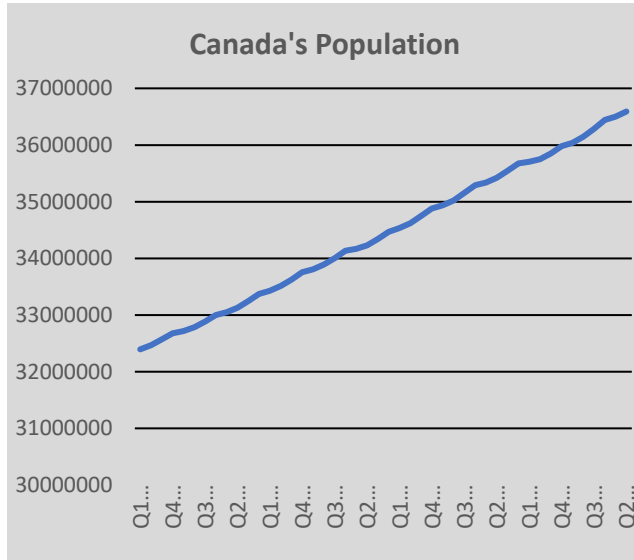


Figure 8: Low variation around trend

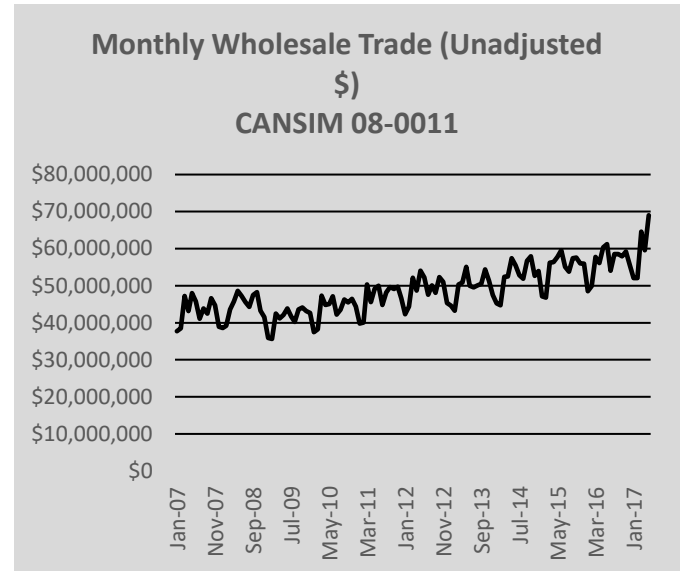


Figure 9: High variation around trend

### 5.3.3. Percentiles and quartiles

The *percentile* is the % of observations that lie below a specific value. For example, if you are in a large class and score in the 75th percentile, that means that 75% of students scored less than you. It is easy to create a percentile; just determine the percentage of each age. The *quartiles* are the three numbers that divide a population into quarters and correspond to the 25%, 50%, and 75% percentiles.

[Mincome Female Heads - Percentile.xlsx](#)

## 6. Summary

Completing this chapter (reading the text, working the examples, and watching the videos) means you can create increasingly complex formulas. Pay attention to =IF and =COUNTIF, as these will be workhorses throughout the course. To understand these formulas, create small examples in Excel and apply the formulas. For example, to understand the difference between the population and sample variance, create a list of numbers and use the two formulas. Then, compute the values from the actual statistical formula (variance = the sum of squared deviations divided by n) and obtain the same value as the Excel formula. Start using Excel as a mathematical “scratch pad” and just play around with examples; this increases familiarity and speed, leading to increased productivity.

Module 3: Quantitative Economics 1- Basic Statistics

Annex: Key Excel function and formulas

Function/formula	Example	Explanation
=IF(criteria,then,else)	=IF(C1=3,1,0)  =IF(C1<>3,1,0)	In the cell where this formula appears, if C1=3, then place the value 1, otherwise place 0.  In the cell where this formula appears, if C1 is not equal to 3, place 1, otherwise place 0.
=COUNT(cell range)	=COUNTIF(A1:F300)	Counts the number of cells with numbers.
=COUNTIF(cell range, criteria)	=COUNTIF(A1:F300,AJ31)	Counts the number of cells in the range A1:F300 that have values equal to the entry in AJ31).
=AVERAGE(cell range)	=AVERAGE(A1:A100)	Finds arithmetic average of the range A1:A100. (Does not count cells with blanks, text, or characters.)
=MEDIAN(cell range)	=MEDIAN(A1:A100)	Finds the median in the range. (Does not count cells with blanks, text, or characters.)
=MODE(cell range)	=MODE(A1:A100)	Finds the mode in the range. (Caution: Multiple entries of the same number must exist in the range, otherwise this will return the error #N/A.)
=GEOMEAN(cell range)	=GEOMEAN(A1:A100)	Finds the geometric mean in the range. (Does not count cells with blanks, text, or characters.)
=MIN(cell range) =MAX(cell range)	=MIN(A1:A100) =MAX(A1:A100)	Computes the minimum and maximum of the range. Takes the difference for the range.
=ABS(cell)	=ABS(A1)	Computes the absolute value of the cell contents.
=VAR.P(cell range) =VAR.S(cell range)	=VAR.P(A1:F300) (Population variance) =VAR.S(A1:F300) (Sample variance)	Computes the population variance (sum of squared deviations divided by n). Computes the sample variance (sum of squared deviations divided by n-1).
=STD.P(cell range) =STD.S(cell range)	=STD.P(A1:F300) =STD.S(A1:F300)	Computes the standard deviation for the population (square root of sum of squared deviations divided by n). Computes the standard deviation for the sample (square root of the sum of squared deviations divided by n-1).
These versions of variance and standard deviation ignore non-numerical values (dates, blanks, characters, text) in the range.		