LSE School of Public Policy

Executive Master of Public Administration

Civil Service and LSE Executive MPP

MATHS AND STATISTICS

PRE-ARRIVAL MODULES

Throughout these modules, you might encounter concepts that are unfamiliar or that you may not understand. *This is fine*. All of the information in these pre-arrival modules will be reviewed during Introduction to Statistics as well as in your core curriculum.

You should study these modules at your own pace, taking care to understand the various concepts given. The four modules are:

Module 1: Building Blocks of Policy
Module 2: Cause and Effect
Module 3: Mathematical Notation, Equations, and Functions
Module 4: Descriptive Statistics and Visualizing Data

Feel free to review the modules you need. Key concepts or terms are outlined in red. Practice problems can be found in a separate pdf, which you can check using the answer key provided.

Module 1: Building Blocks of Policy

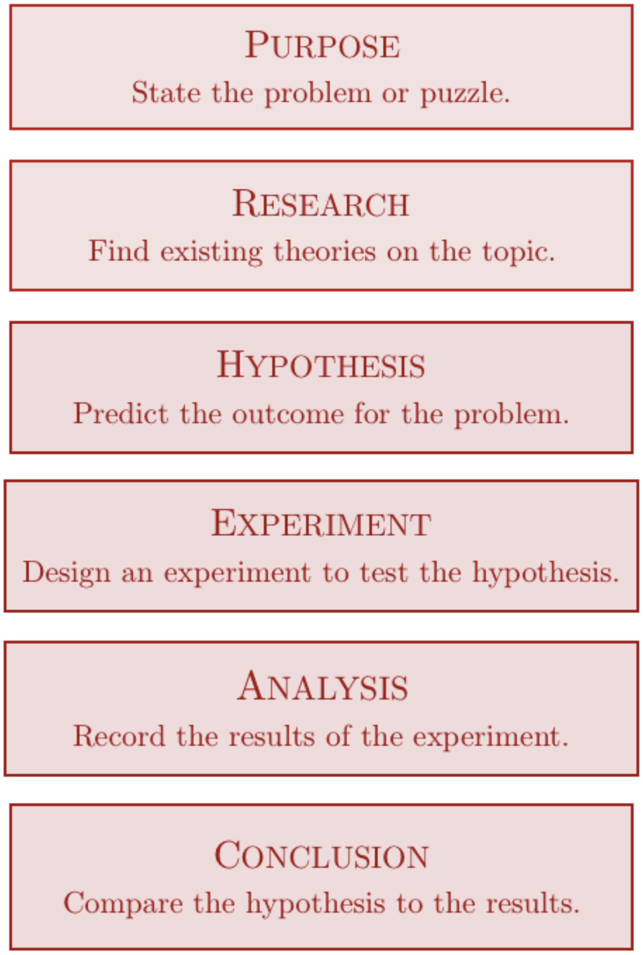
All fields in the social sciences rely on the *scientific method,* or the systematic and logical method of problem solving using experiments that originated in the natural and physical sciences.

The diagram on the right presents one way to think about the scientific method.

There are a number of steps in this process — from observation and research question, to theories and models, to hypothesis testing and ultimate research design. All rigorous research follows some version of these steps.

This module reviews these elements, and provides a number of useful definitions for modern public policy research.

The Scientific Method



Purpose
State the problem or puzzle.

Research
Find existing theories on the topic.

Hypothesis
Predict the outcome for the problem.

Experiment
Design an experiment to test the hypothesis.

Analysis
Record the results of the experiment.

Conclusion
Compare the hypothesis to the results.

1.1 Research Questions, Theories, and Hypotheses

Any academic analysis in the social sciences typically begins with a question. The *research question* is the problem or puzzle to be investigated in a study. For example:

- How do we prevent corruption in low information environments with weak institutions?
- Why have we seen the rapid rise of populism in Europe?
- How do we increase the number of women or minorities in politics?
- Would we get better policy if we had better politicians or more informed voters?
- Can federal systems prevent ethnic conflict?

We always start with a question, and then we develop models, and then theories, to come up with an answer which we can then test by collecting evidence.

A *theory* is a set of logically consistent statements that tell us why the things that we observe occur. It is essentially an explanation, that is supported by preexisting evidence. It allows us to understand how and why already observed phenomena occur, and helps us predict as yet unobserved relationships.

A good theory is logically consistent, and is specified in a way that has testable, empirical predictions that could be falsified (or proven wrong). A poor theory typically cannot be proven wrong using empirical evidence (for example, "ghosts exist") or is underdeveloped such that it almost restates the observation to be explained (for example, "developing countries are poor").

Relatedly, a research *hypothesis* (plural, hypotheses) is a prediction or assumption about what will be found when the data are collected and analyzed. This is constructed before the research study or any data collection begins, though a hypothesis is often formulated with existing theories and past research in mind. It is typically more specific, and usually indicates the direction of the causal effect.

For example, a testable hypothesis would be that "*Individuals with a college education will earn a higher salary than individuals without a college education.*" This is a clear prediction that can be empirically analyzed, and we know the hypothesized direction here is positive — as the number of years of education increases, salary also should increase.[1]

Often you will hear the term hypothesis testing, which is the goal of all research using statistical analysis. A research study is typically described as trying to adjudicate between a "null hypothesis" and an "alternative hypothesis."

The *null hypothesis* (indicated by H_0), states is that there is no effect.

The *alternative hypothesis* (indicated by H_A) is created by the researcher and describes the expected relationship, usually building on preexisting theories or empirical patterns.

[1] For more on the difference between a theory and a hypotheses: https://www.merriam-webster.com/words-at-play/difference-between-hypothesis-and-theory-usage

The table below provides an example of a research question, theory, and hypotheses using years of education and future salary.

Research Question	Why does A happen? What is the relationship between A and B?	To what extent does a college education impact future salary?
Theory	A affects B	Employers value advanced skills, so college education has a positive affect on future salary.
Hypothesis (H_A)	Increases in A result in an increase in B.	People with an undergraduate degree will have a higher salary than those with just a high school degree.
Null Nypothesis (H_0)	There is no relationship between A and B.	There is no relationship between college education and salary.

1.2 Methods and Hypothesis Testing

Once we have established the research question, theories, and corresponding hypotheses; it is time to take these to data. This is how we test a hypotheses — we collect evidence to prove that the hypothesis is wrong or right.

The body of research methods used in an academic discipline is called its *methodology*. There are a wide variety of research methods used to approach problems in the social sciences. Accordingly, a researcher needs to be very thoughtful about choosing their methodology. The research question, and ultimate design of the study, should dictate the methods used.

Importantly, the study of public policy is naturally interdisciplinary, pulling from economics, political science, public administration, sociology, and a wide variety of other disciplines as needed. Though the research methods can vary, often we distinguish between two categories: "qualitative" or "quantitative" methods.

These are broad and by no means mutually exclusive categories; often good research involves an element of both. Further, qualitative insight is needed to make sense of quantitative coding or analysis. Still, this terminology is common, and useful to review.

Qualitative research focuses primarily on gathering non-numerical, descriptive information on the attributes or characteristics of the objects in the study.

> These can include case studies and detailed contextual research, process tracing, comparative analysis, ethnographic research, biographical or diary accounts, or unstructured interviews. Any analysis is descriptive and non-statistical, and usually features a small and unrepresentative sample. Qualitative data typically investigates *why* things happen. While not involving "hard data," qualitative methods require a strong theoretical grounding and structured framework of comparison in order to accurately interpret such nuanced evidence.

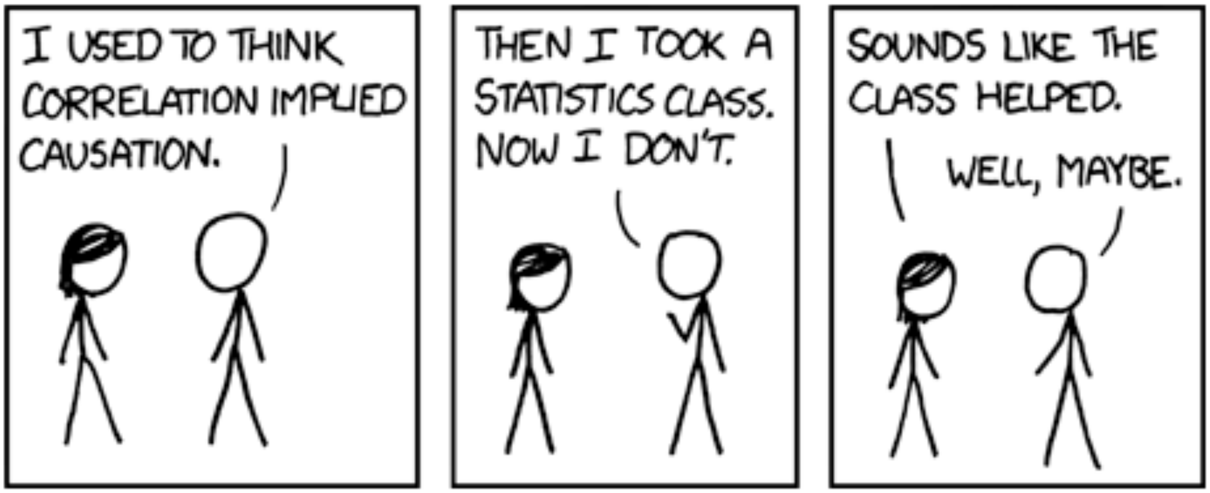
Quantitative research gathers and analyzes numerical data.

> This could be done using mathematical methods, statistical analysis, quantitative text analysis, or a variety of other numerical techniques. Data can be collected from preexisting sources, or generated using experiments or surveys, but the sample size is typically large and representative. Quantitative data typically is focus on determine "*how much*" or "*how many*."
>
> There are two main types of quantitative data:
>
> - observational data, or data collected from the world around us
> - experimental data, or data generated under conditions controlled by the researcher

Module 2: Cause and Effect

One of the primary goals of public policy analysis is to study the effect of a policy or set of policies on an outcome of interest. For example, we want to identify the effect of new textbooks on student test scores, a new health policy on mortality rates, or the effects of interest rates on economic growth. To do so, we have to systematically describe and measure the many relationships between a policy and its resulting outcomes.

To tackle these problems, we often structure our thinking in a scientific way, using the concepts of "cause" and "effect" that come directly from experimental research. Further, we typically express these relationships using mathematical notation. This module introduces the idea of causality and ways this is expressed using mathematical notation (which will continue in Module 3), and highlights the main issues that interfere with our ability to measure cause and effect.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

https://xkcd.com/552/

2.1 Introductory Notation: Variables and Path Diagrams

A causal relationship is nearly always expressed using variables.

A *variable* represents a concept or entity that can be measured quantitatively, or in other words, that can represented by a number. A variable can take on a range of different values. For example, a variable could measure the number of houses in a city, the income of a person, or the number of international treaties signed on climate change (for more on measurement of variables, see Module 4).

Importantly, variables are single symbols, represented typically by letters of the alphabet (A, b, c) or Greek letters ($\boldsymbol{\beta}$, λ, $\boldsymbol{\phi}$). The two most common variables, and also used to express the idea of cause and effect, are X and Y.

The letter X represents the *independent variable*, or "cause," in the relationship of interest. In an experiment, this is the variable the researcher manipulates, and is assumed to have a direct effect on the outcome (Y). In a non-experimental study, or observational study, this variable simple represents *potential* causes of the outcome of interest.

The letter Y represents the *dependent variable,* which is the ultimate "effect" or "outcome." In an experiment, this variable is expected to change depending on the manipulated values of the independent variable (X), and it is what the researcher measures at the conclusion of the study. It is called the dependent variable because its values "depend" on changes in the independent variable (X).

Often we describe theoretical concepts relating to cause and effect using path diagrams. A path diagram is a picture in which arrows are drawn from (perceived) causes to their (perceived) effects.

You will see such conceptual diagrams, where variables connected by an arrow, such as:

$$X \longrightarrow Y$$

This can be translated as X causes Y, or more precisely, X causes a change in Y. Sometimes you also may see a plus (+) or minus (—) sign above the arrow, indicating the direction of the effect (i.e., the independent variable has a positive or negative effect on the dependent variable).

For example, in Module 1 we discussed the relationship between education and future salary. In that policy problem, the independent variable (X) or "cause" would be college education, and the dependent variable (Y) or "effect" would be increased salary.

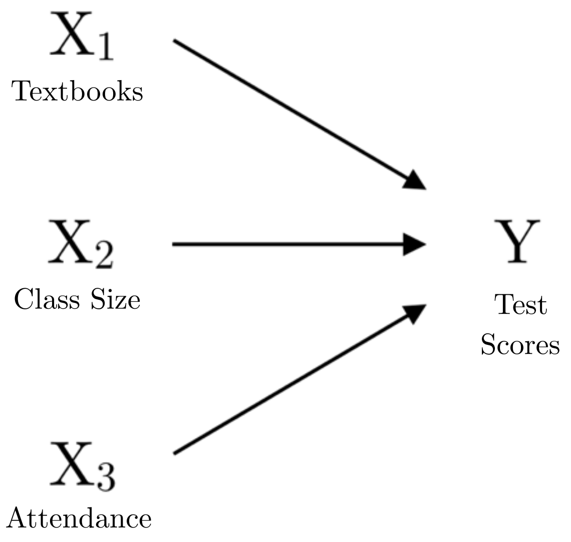
This is how we typically interpret these types of path diagrams:

X	$\longrightarrow$	Y
Independent variable	causes (a change in)	Dependent variable

Examples:

Education	causes (a change in)	Salary
Interest Rates	cause (a change in)	Economic Growth
New Textbooks	cause (a change in)	Test Scores

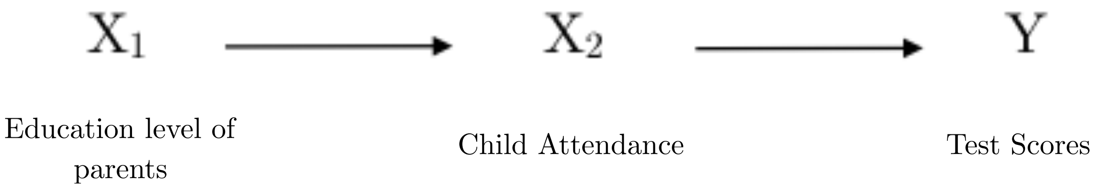
Path diagrams also help us to visualize and logically consider multiple explanations for an outcome. Consider a different policy problem: do new textbooks (X) increase test scores of students (Y)? Here, we might think that our outcome of interest — test scores — is caused by a number of factors.



In the path diagram to the left, multiple independent variables are drawn, but are distinguished by a subscript — i.e. X_2, X_3 — indicating they are a separate independent variables, or separate causes.

This diagram helps demonstrate that there are multiple causes of high test scores (new textbooks, smaller class size, student attendance) that we may want to separately consider.

We also might think that one variable works indirectly, through another variable. For example, educated parents might ensure their child attends school, which then translates into better test scores. The path diagram below demonstrates this relationship.



It is also very important to note that these path diagrams are NOT equations. Diagrams simply help us visualize the relationships between variables, to allow us to develop theories and testable hypotheses. In contrast, equations have equal signs (=) and can be manipulated and solved. Often the letters or symbols chosen to represent these variables will be identical in the corresponding statistical equation. Equations will be covered in Module 3.

2.2 Causal Inference

The goal of any research in public policy, political science, or economics is to infer the causal effect of one variable on another. Yet even if we have theories about the relationship between two variables, it is often challenging to prove causality using data.

The field of *causal inference* aims to identify the effect of X on Y, using advanced econometric techniques. In policymaking, this could be the effect of exposure to a particular treatment or program; or as in our previous example, the effect of education on salary.

The gold standard in policy research is to conduct a controlled experiment in which treatments (programs) are allocated at random, and many of your EMPA courses will focus on providing you the specifics of this skill set. Before that, however, it is useful to review the basic challenges to causal inference, using the introductory notation presented above.

2.2.1 Correlation Does Not Imply Causation

Often academic research begins by looking at an association between two variables. Yet while simply finding an association might be suggestive of a causal effect, it is not guaranteed — or as the saying goes, "correlation does not imply causation."

A correlation is a measure of association, and is defined as the degree of relationship between two variables. Said in another way, to measures how strongly they are linked together. If there is no observed relationship between the two variables, there is no correlation.

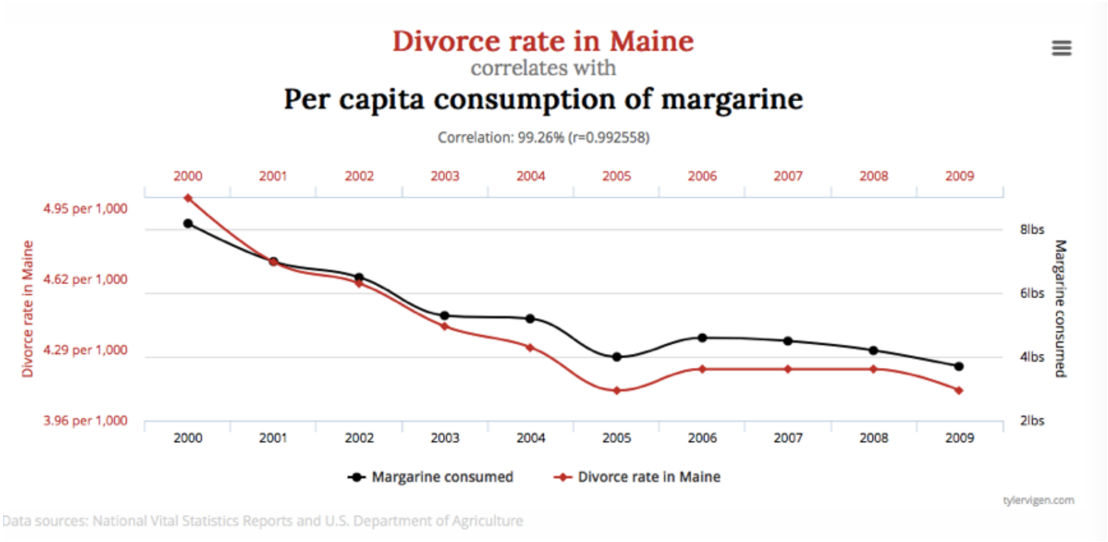
However, sometimes we find a correlation (relationship) between two variables, but this relationship isn't causal; this is discussed at the end of this section. More on how to measure and visualize correlations using data can be found in Module 4.

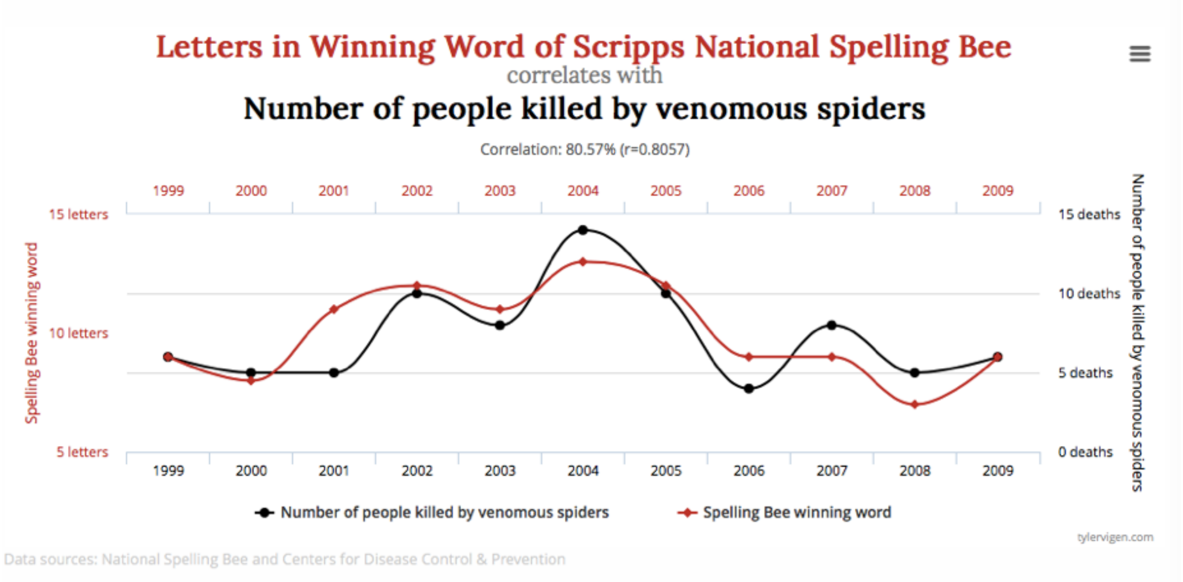
A positive correlation is when the values of each variable move together, such that their values increase or decrease together. For example, height and weight are positively correlated: as height increases, typically weight also increases. Note that two variables can be positively correlated, but have decreasing values, in that as one variable decreases the other also decreases (for example, as attendance at school decreases, then academic achievement also decreases). This is because correlation measures the strength of the relationship between the two variables.

A negative correlation is when the two variables don't move together, or as the values of one variable increase the values of the other variable decrease. For example, the price of the London Congestion charge and the number of cars in London are negatively correlated: as the price of the toll for cars driving in London *increases*, it *decreases* the number of cars on the road.

Establishing a correlation between two variables is relatively straightforward — we can find often find empirical data to suggest that two things move together. But the next two charts demonstrate that a correlation is not always indicative of a causal relationship. Or, said in another way, just because two things are correlated doesn't mean they have any substantial relationship to each other.

These charts present actual empirical data that demonstrates strong correlations — except between two random variables. The first demonstrates that the divorce rate in Maine is almost perfectly correlated with margarine consumption, and the second shows a strong correlation between the number of letters in the words in a spelling competition and deaths by venomous spiders. Clearly, these variables — while highly correlated — have no causal relationship to each other. No one believes that spelling competitions cause death by spiders, or vice versa. These are called spurious correlations.





"Spurious Correlations" (http://www.tylervigen.com/spurious-correlations)

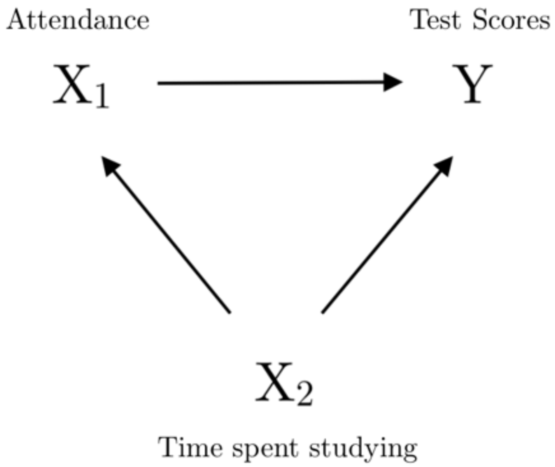
In general, it is good to keep in mind "correlation does not imply causation." However, correlation remains a good starting point for analysis, particularly if the correlation can be grounded in existing theories. Further, a correlation might imply causation, but the causal effect may operate differently. For example taking cough medication is positively correlated with frequency of coughing but hopefully has a negative causal effect on coughing. Before conducting any econometric analysis, it is worth spending time to logically consider all relationships.

2.2.2 Omitted Variables

Omitted variable bias is a statistical term, that describes a situation in which a model or explanation leaves out an important independent variable (also called a confounder). However, it also demonstrates a larger issue with either our theory or research design.

For example, say we wanted to explore the relationship between school attendance and test scores. We can diagram this relationship, and if we looked at data we would see it is positively correlated — as attendance improves, a student's test scores will also improve. This may lead us to conclude that increasing attendance has a causal effect on student performance.

However, there might be other important causal factors we are missing. The time a student spends studying might be equally, or more important, than attendance. Graphing these relationships in a path diagram is a good way to logically think through causal mechanisms; here, X_2 is the omitted variable.



In this example, study time is the omitted variable — it has an effect on both our independent variable and dependent variable of interest. If we leave out study time, we omit an important causal mechanism; as as you will learn in your EMPA classes, it biases our results and sometimes leads to incorrect conclusions.

Often case specific knowledge is needed to determine missing causal variables. The adoption of compulsory voting practices around the world is a good example of this issue.[2] A number of countries, such as Australia or Brazil, have policies in place that make it mandatory for citizens to vote.

If we consider the adoption of compulsory voting policy the treatment, which is our independent variable (X), we might be interested in a number of potential outcomes, or dependent variables (Y). Observational data shows strong a correlation between compulsory voting and political participation, voter information, and leftist parties; depending on the research design, we may conclude that compulsory voting either causes or is at least associated with these outcomes.

However, it is also true that the types of countries that adopt compulsory voting policies might be systematically different that countries who do not. For example, countries with informed voters and cultural norms of egalitarianism are more likely to adopt compulsory voting. This reflects their preexisting values, which could predict both the adoption of compulsory voting rules and the empirical outcomes (such as support for leftist parties) we see in the data. Without considering such omitted variables, we might incorrectly attribute outcomes to the adoption of a compulsory voting policy.

2.2.3 Reverse causality

Reverse causality is another threat to inference. This is when it is possible that either X causes Y, or Y causes X. It is up to the researcher to first substantively analyze to what extent reverse causality could be a problem. This is sometimes straightforward, and sometimes more difficult.

[2] For more on this topic, see Hangartner et al. (2016). “Does Compulsory Voting Increase Support for Leftist Policy?” *American Journal of Political Science*, Vol. 60, No. 3, July 2016, Pp. 752–767.

The next table presents two different examples of observed correlations. Are both these relationships theoretically possible?

Observed Correlation	X Independent Variable	→	Y Dependent Variable	Possible?
1. The use of umbrellas and rainfall are positively correlated.	Umbrellas	→	Rainfall	No, umbrellas do not cause it to rain.
	Rainfall	→	Umbrellas	Yes, rainfall causes an increase in umbrella usage.
2. Higher levels of GDP and democratic regimes are positively correlated.	GDP	→	Democratic Regime	Potentially — higher GDP makes it more likely a country will become or remain democratic.
	Democratic Regime	→	GDP	Potentially — democratic institutions can promote GDP growth.

In the first example, it is easy to deduce that the causal relationship can only work in one direction. Simple laws of nature tell us that using umbrellas can't cause it to rain, so the causal relationship should be conceptualized with rain as the independent variable (X) and umbrella usage as the dependent variable (Y).

In the second example, it is much more difficult. There are theoretical arguments to be made that increasing GDP both causes democratic regimes, and that democratic regimes cause an increase in GDP. Therefore in this example, there is a problem with reverse causality, which must be addressed in the researcher's study. (This is an ongoing debate in political science, which you will learn more about during the EMPA program).

2.3 A Note on Causal Terminology

As a result, most academic papers are relatively precise about when they claim a set of findings is causal. Papers using a randomized controlled trial, for example, will write that their results are causal ("A smaller class size *causes* an increase of 10 points on annual test scores"). Papers using observational data (data obtained without the use of an experiment) try to control or include all the variables they can, but typically won't, and shouldn't, be this bold. Instead, these studies will describe their independent variable as being *associated.*

Colloquially, you might also hear this problem discussed using the term "identification." For example, "This study has problems with its identification strategy." This simply means that the study cannot confidently claim causality. You will learn more about this during your first year studies.

Associational Terms	Causal Terms
correlation	randomization
regression	intervention
likelihood	instrumental variables
"is associated with"	exogoeneity
"controlling for"	identification

For an interesting and current example on causal language, see: "The correlation-causation two-step, police shootings edition" at https://scatter.wordpress.com/2017/03/14/the-correlation-causation-two-step-police-shootings-edition/

Module 3: Mathematical Notation, Equations, and Functions

This module reviews basic concepts relating to reading mathematical notation, understanding equations and functions, and manipulating or solving equations.

You will encounter these concepts frequently in your EMPA courses, and you will also need to be able to engage with the academic literature on various subjects. Both political science and economics papers often use a lot of math, so you will need to know enough mathematical concepts to be able to intelligently read these materials. Further, as your studies progress, you will find that the language of mathematics often helps us present complex policy problems in a more simple way, that better allows us to design a credible policy solution.

This module begins with mathematical notation, and presents commonly used symbols, letters, and abbreviations in mathematics and statistics. It then reviews the difference between equations and functions, as well as how to solve or manipulate mathematical terms. Some of this content is just meant as a quick review; don't worry if you don't understand all of it.[3]

It also may seem that there is no connection between solving an equation for "x" and developing policies or programs. Yet equations are the next step in presenting how one variable relates to another, and more importantly they provide the foundations for statistics — which allows us to take our hypothesized causal relationship between X and Y to data that we collect in the real world. Statistics will be covered in your first year courses, so no need to worry about this now. But for those of you who haven't used equations since college, this is a useful section to spend time on.

[3] Also, a very useful resource is *A Mathematics Course for Political and Social Research.* (2013). Will H Moore and David Siegel. Princeton University Press.

3.1. Notation

It is often necessary to express the relationship between X and Y more formally. This requires a basic understanding of mathematical notation, as well as concepts such as functions and equations.

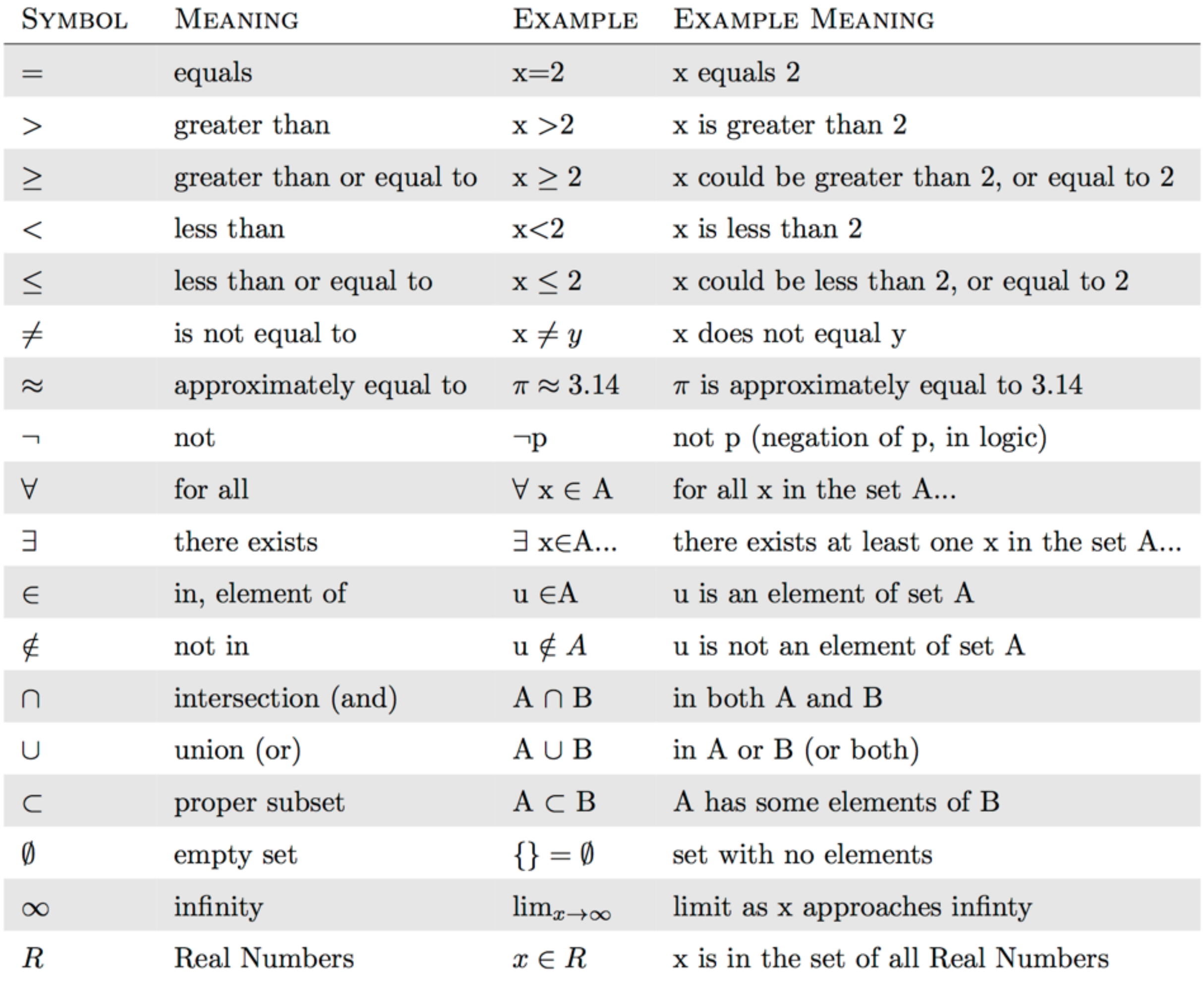
Tables 3.1 and 3.2 presents some commonly used statistical abbreviations, as well as mathematical notation. Greek letters are also often used in statistics and economics courses; Table 3.3 presents the set of Greek letters, uppercase and lowercase, and their pronunciation. We will explore more about what these mean during the course of the module. Note, this is not a comprehensive list (for more information, see the resource list at the end of this module).

By convention, some letters or symbols consistently represent the same information; however, variables are just placeholders and can be chosen by researchers. Each study will specify what symbols represent which information, but these may vary across studies. Some of this notation you will see very often; other symbols you may not see as much. However, it is good to skim them all, just in case.

Table 3.1: Useful Abbreviations

SD	Standard Deviation
SE	Standard Error
CI	Confidence Interval
P(A)	the probability of event A
P(B |A)	probability of B given A
H_0	Null Hypothesis
H_A	Alternative Hypothesis
s.t.	“such that”
w.l.o.g.	“without loss of generality”
iff	“If and only if”

Table 3.2: Common Mathematical Symbols



Symbol	Meaning	Example	Example Meaning
$=$	equals	x=2	x equals 2
$>$	greater than	x >2	x is greater than 2
$\geq$	greater than or equal to	x $\geq$ 2	x could be greater than 2, or equal to 2
$<$	less than	x<2	x is less than 2
$\leq$	less than or equal to	x $\leq$ 2	x could be less than 2, or equal to 2
$\neq$	is not equal to	x $\neq y$	x does not equal y
$\approx$	approximately equal to	$\pi \approx 3.14$	π is approximately equal to 3.14
$\neg$	not	$\neg$p	not p (negation of p, in logic)
$\forall$	for all	$\forall$ x $\in$ A	for all x in the set A...
$\exists$	there exists	$\exists$ x$\in$A...	there exists at least one x in the set A...
$\in$	in, element of	u $\in$A	u is an element of set A
$\notin$	not in	u $\notin A$	u is not an element of set A
$\cap$	intersection (and)	A $\cap$ B	in both A and B
$\cup$	union (or)	A $\cup$ B	in A or B (or both)
$\subset$	proper subset	A $\subset$ B	A has some elements of B
$\emptyset$	empty set	$\{\} = \emptyset$	set with no elements
∞	infinity	$\lim_{x\to\infty}$	limit as x approaches infinty
R	Real Numbers	$x \in R$	x is in the set of all Real Numbers

More generally, mathematical notation provides a shorthand for expressing complex ideas. While some phrases may look complicated initially, especially to those who aren't used to "reading math," you will find they are much more straightforward to interpret after a review of frequently used symbols.

Table 3.3: Greek Letters

Lowercase	Uppercase	Name (Pronunciation)
α	A	alpha (AL-fuh)
β	B	beta (BAY-tuh)
γ	Γ	gamma (GAM-uh)
δ	Δ	delta (DEL-tuh)
ϵ	E	epsilon (EP-sil-on)
ζ	Z	zeta (ZAY-tuh)
η	E	eta (AY-tuh)
θ	Θ	theta (THAY-tuh)
ι	I	iota (eye-OH-tuh)
κ	K	kappa (KAP-uh)
λ	Λ	lambda (LAM-duh)
μ	M	mu (MYOO)
ν	N	nu (NOO)
π	Π	pi (PIE)
ρ	R	rho (ROW)
σ	Σ	sigma (SIG-muh)
τ	Y	tau (TAU)
υ	Υ	upsilon (OOP-si-LON)
ϕ	Φ	phi (FEE)
χ	X	chi (K-EYE)
ψ	Ψ	psi (SIGH)
ω	Ω	omega (oh-MAY-guh)

3.2. Equations and Functions

An *equation* is a declaration that two expressions are equal to each other. Most often, it is a mathematical expression containing variables of unknown value. Here are some examples of equations:

$2^2 = 4$
$y = mx + b$ (linear equation)
$x^2 + bx + c = 0$ (quadratic equation)
$x^2 = 4$

In the last equation, $x^2 = 4$ is true when $x = 2$ or when $x = -2$, but false for all other values of x. In this, note a linear equation can have multiple solutions.

A function is also a relation between two or more variables, but is much more complex. It describes a mathematical relationship in which the values of a single dependent variable are determined by the values of one or more independent variables. A linear function is essentially a machine: you input one number, and a rule is used to output a corresponding number. As a result, every value of x can only have one value of y.

For example, this is a linear function:

$$f(x) = x + 1$$

This also is equivalent to:

$$y = x + 1$$

A function always has three parts:

1) the input, x, also called the independent variable
2) the relationship
3) the output, f(x), also called “y” or the dependent variable

The classic way of writing a function is with "f(x) = ... ", where x denotes the input and f(x) denotes the output. Note that the function is f, and f(x) is simply the value of the function for a given value of x.

For example, if you input 4 as x, your output is 5.

$f(x) = x+1$

$f(4) = 4+1$

$f(4) = 5$

Note that a function assigns only one value of y to any x. However, any given value of y can correspond to multiple values of x.

For example:

$$y = x^2$$

If the value of y is 4, the value of x could equal either 2 or -2.

3.3 Algebra Review

In order to review how to manipulate and solve equations, it is useful to recall rules from algebra.

3.3.1 Order of Operations

The order of operations is the order in which expressions are evaluated.

1. Brackets and parentheses.
2. Exponents and roots.
3. Multiplication and division, from left to right.
4. Addition and subtraction, from left to right.

For example, evaluate the following expression.

$$\begin{aligned}
&16 - 3(5 - 3)^2 \div 2 \\
&= 16 - 3(2)^2 \div 2 && \textit{Parentheses first} \\
&= 16 - 3(4) \div 2 && \textit{Exponents} \\
&= 16 - 12 \div 2 && \textit{Division first} \\
&= 16 - 6 && \textit{Then subtraction} \\
&= 10
\end{aligned}$$

3.3.2 Basic Algebra

Much of algebra involves manipulating complex equations, in order to simplify and then solve them.

When simplifying terms, one key rule to remember is that you can only add or subtract terms that have the exact same variable.

For example:

$x + x$ can be simplified to $2x$

$5xy + 5xy$ can be simplified to $10xy$

$x + x^2$ <u>cannot</u> be simplified or combined (it is <u>not</u> x^3 or $2x^2$)

However, you can multiply or divide across terms. Note that the multiplication symbol can either be expressed using the symbols × or · but both mean that a number or variable is "times" another (eg, $x{\cdot}x$ is pronounced "x times x").

$x{\cdot}y$ can be simplified to xy

$x{\cdot}x$ can be simplified to x^2

$8xy/2x$ can be simplified to $4y$

When manipulating equations, you can add, subtract, multiply or divide by any number or term, as long as you do the same thing to each side.

For example, solve for b in the equation $y = mx+b$.

$y = mx+b$

$y - mx = b - mx$ *Subtract mx from each side*

$y - mx = b$

$b = y - mx$ *Arranging so b is on the left*

Equations demonstrate the relationship between multiple variables, and can be solved for specific values.

To take a very simple example, imagine a school has a yearly budget of $1,000, and wants to know how many new textbooks it can buy with that budget. Textbooks cost $50 each.

One way to write this using an equation is as follows: 50x = 1,000

Here, x represents the (unknown) number of textbooks a school can buy, and is multiplied by 50 which is the price of each textbook. The school budget is 1,000, which is the maximum amount it can spend.

We want to solve this for x, to find the number of textbooks the school can buy given its budget.

$$\begin{aligned} 50x &= 1000 \\ x &= \frac{1000}{50} \\ x &= 20 \end{aligned}$$

Here, we can quickly solve and see that the school can buy 20 textbooks.

3.3.3 Rules and Identities

Here are a number of examples of useful rules and identities in algebra that might come in handy. No need to memorize, just review them for clarity. These are:

Rule/Identity	Example using numbers
$a+b=b+a$	$1 + 2 = 2 + 1$
$(a+b)=(b+a)$	$(1+2) = (2+1)$
$a+0=a$	$1+0 = 1$
$a+(-a)=0$	$1 + (-1) = 0$
$ab = ba$	$1\cdot 2 = 2\cdot 1$
$(ab)c = a(bc)$	$(1\cdot 2)3 = 1(2\cdot 3)$
$1\cdot b=b$	$1\cdot 2=2$
$(-a)b = a(-b)$	$(-1)2 = 1(-2)$
$(-a)(-b) = ab$	$(-1)(-2) = 1\cdot 2$
$-(a+b)=-a-b$	$-(1+2)= -1 -2$
$a(b+c)=ab+ac$	$1(2+3) = 1\cdot 2 + 1\cdot 3$

3.3.4 Multiplying and Factoring Algebraic Terms

You also may need to expand or multiply terms. When multiplying terms in parentheses, take each term enclosed in the first set of parentheses and multiply it against the terms in the second:

$$(a + b)(c + d)$$

$$= a(c + d) + b(c + d)$$

$$= ac + ad + bc + bd$$

Or in the case with both negative or minus signs (-) and positive or plus (+) signs:

$(a - b)(c + d)$

$= a(c + d) - b(c + d)$

$= ac + ad - bc - bd$

Other useful identities are:

$(a - b)(a + b) = a^2 - b^2$

$(a + b)^2 = a^2 + 2ab + b^2$

$(a - b)^2 = a^2 - 2ab + b^2$

Factoring is the opposite of expansion, and is used to solve quadratic equations. It involves expressing the equation as a product of two roots. This is often a matter of trial and error. Other methods include completing the square, or using the quadratic formula.

For example, factoring $x^2 + 5x + 6$ results in a solution of $(x + 3)(x + 2)$.

3.3.5 Percentage versus Percentage Point

For making good public policy, it is very important to understand the difference between a percentage and a percentage point. Often percentages are accidentally misused by people who don't understand this difference, or even intentionally misused. For example, we might be concerned if in the past year, it was reported that the number of schools closed due to poor performance increased by 100%. This sounds concerning. However, if the number of failed schools has risen from 1 to 2, in a district with over 100 schools, then this is still a relatively small substantive change.

In another example, we might think that a new environmental policy increased the frequency with which citizens recycle. However, if you see the headline that “Recycling increased from 10% to 12%,” what does this mean? Did it rise by 20% (12/10=1.2, so .2 or 20% change) or did it rise by 2%? Here it is important to know the difference between a percentage change and a percentage point change.

A *percentage* is a number or ratio expressed as a fraction of 100.

For example, if 30% of voters support the Green party, then 30 people out of every 100 support the Green party. To change a fraction to a percentage, divide the numerator by the denominator and multiply the resulting decimal by 100.

$$\frac{30}{100} = .3 * 100 = 30\%$$

Sometimes you will need to calculate a percentage when the numbers are not easily converted to 100. Here, you divide the numbers by each other and multiple by 100. For example, 66 is what percentage of 94?

$$\frac{66}{94} = .7 * 100 = 70\%$$

A *percentage change* (increase or decrease) shows the ratio of the change in two numbers. First, calculate the change, by subtracting the original number from the new number; then divide the change by the original number to get a decimal. Multiple this by 100 to get the percentage change.

$$\text{Percentage Change} = \frac{New\ Number - Original\ Number}{Original\ Number} * 100$$

For example, say the number of recycling facilities increased from 5 to 7. What is the percentage change?

$$\frac{(7-5)}{5} * 100 = \frac{2}{5} * 100 = .4 * 100 = 40\%$$

Here, the difference in facilities is 7-5, which is 2. Then 2 divided by 5 is .4, and multiplied by 100 is 40%. The number of recycling facilities increased by 40%.

A *percentage point* change is just the simple numerical difference between two percentages, used when adding or subtracting one percentage from another. It is a difference, not a ratio. Percentage points are useful because they show the change in the quantities of interest with *respect to its previous values.*

Back to our recycling example, if "Recycling increased from 10% to 12%," then this means both:

a) Recycling increased by 20% (2/10=.2, so 20% change)

b) Recycling increased by 2 percentage points

To prevent ambiguity, it is often useful to use both in a policy report. For example, you could say "Recycling increased by 2 percentage points after the new policy, which resulted in a 20% increase in individuals recycling."

3.3.6 Rounding

Numbers can consist of many digits, and rounding is a way to simplify a complex number, often for ease of presentation.

First choose the place you want to round to. Numbers can be rounded to the nearest ten, the nearest hundred, the nearest thousand, etc. Similarly, in the case of decimal numbers (a number with a decimal point) sometimes you might be given the equivalent instructions of "round to one decimal" or "round to two decimals."

These are the rounding guidelines:

1. If the place you want round to is followed by 5 or higher (5, 6, 7, 8, or 9) then increase the final digit by 1 and drop the rest.

For example:

- 68 rounded to the nearest ten is 70
- 2,899 rounded to the nearest hundred is 2,900

2. If the place you want round to is followed by 4 or lower (4, 3, 2, 1, 0) then round the final digit down and drop the rest.

For example:

- 63 rounded to the nearest ten is 60
- 1,209 rounded to the nearest hundred is 1,200

Take the number 3,717.

- 3,717 rounded to the nearest ten is 3,720
- 3,717 rounded to the nearest hundred is 3,700
- 3,717 rounded to the nearest thousand is 4,000

Fractions are rounded in the same way (here, the decimal place rounded to is underlined).

- $1.\underline{7}199$ rounded to the nearest tenth is 1.7
- $0.1\underline{5}501$ rounded to the hundredths place is 0.16
- $5.95\underline{9}2$ rounded to the nearest thousandth is 5.959

Here is an example that rounds up or rounds down, depending on the place we want to round to.

- If only one decimal is to be kept, then 2.416 becomes 2.4.
- If only two decimals are to be kept, then then 2.416 becomes 2.42.

3.4 What is a model?

Across the social sciences, researchers often organize the many relationships of cause and effect using a model. A *model* is a *simplified* description of reality, designed to yield testable hypotheses about cause and effect relationships.

A model is simplified version of the world, so inevitably, most models leave lots of things out. However, we gain the ability to precisely describe and test the phenomenon of interest, in order to gain leverage on understanding some specific aspect of our research question.

Some models are useful to visually express information in an abstract way, while others are much more specific and outline the precise hypothesized relationship between variables. There are also different types of models; here we will discuss i) visual models, ii) mathematical or theoretic models, and iii) empirical models.

3.4.1 Visual models

Visual models are pictures or diagrams to present general concepts or a set of information. Any type of path diagram, or flow chart, is a visual model. While not exhaustive, they are sometimes very useful.

For example, the map of the London tube system is a visual model — it presents information about how the stops relate to each other, in an easy to understand way.

However, it is simplified — the model does not indicate the accurate distance between the stations, for example, causing no end of confusion to tourists.

Paddington
Edgware Road
Baker Street
Portland Street
Euston
Angel
Edgware Road
Marylebone
Warren Street
Euston Square
Farringdon
Old Street
Bethnal Green
Regent's Park
Russell Square
Barbican
Liverpool Street
Bayswater
Notting Hill Gate
Lancaster Gate
Bond Street
Oxford Circus
Goodge Street
Chancery Lane
Moorgate
Holborn
Queensway
Marble Arch
Tottenham Court Road
St. Paul's
Bank
Aldgate East
Covent Garden
Aldgate
High Street Kensington
Hyde Park Corner
Green Park
Leicester Square
Mansion House
Cannon Street
Piccadilly Circus
Knightsbridge
Monument
Tower Hill
Tower Gateway
Charing Cross
Blackfriars
Gloucester Road
Sloane Square
St. James's Park
Temple
River Thames
London Bridge
Earl's Court
South Kensington
Victoria
Westminster
Embankment
Bermondsey
Canada Water

3.4.2 Mathematical and Theoretical Models

Theoretic models aim to derive implications about behavior, under the assumptions that actors are rational and maximizing specific objectives subject to varying constraints.

One common type that you will see in the EMPA program is a game theoretic model. Game theory is the formal study of strategic decision making, where two or more players must make choices that potentially affect the interests of the other players.

For example, the “Prisoner’s Dilemma” is a classic game in game theory. Two members of a criminal gang are arrested by the police, who don’t have enough evidence to convict them of a serious crime unless at least one of them confesses. Each are separately interrogated, and each criminal can either stay quiet or confess. If both stay quiet, they get a minor sentence and one month in jail; if both confess, the police have enough evidence to send them both to jail for one year. If one criminal confesses and the other stays quiet, the first criminal goes free and the second gets a harsher sentence of two years — one for the crime, and one year more for obstructing justice.

Even though it would be in each criminal’s best interest to stay quiet (action “C”, for cooperate), each criminal thinks the other will confess, and the structure of the game counterintuitively incentivizes each to confess (action “D”, for defect).

A game theoretical model describes the players, their strategies, and their payoffs from taking various actions. Here are a few ways to model this interaction; this is a normal form game showing the payoffs for each action for each player.

		Player 2	
		C	D
Player 1	C	2, 2	0, 3
	D	3, 0	1, 1

Next we have a set of inequalities that demonstrates the utility (u_i) of each combination for each player. u_1 represents the utility for player 1, and similarly u_2 represents the utility for player 2; each possible action by both players is listed. The second shows that, for player 1, the utility of him defecting if the other stays quiet (D for player 1, C for player 2) is greater than the utility of both players staying quiet (C, C), which is greater than both players confessing (D, D), which is better than the player 1 confessing while player 2 stays quiet (C, D).

This is another way of expressing the strategic tradeoff between taking one action over another.

For player 1:

$$u_1(D,C) > u_1(C,C) > u_1(D,D) > u_1(C,D)$$

For player 2:

$$u_2(C,D) > u_2(C,C) > u_2(D,D) > u_1(D,C)$$

Mathematical models are simply the representation of concepts or behavior using mathematical language. Often they are founded upon an equation, or set of equations. Such models are typically solved to determine the answer; or sometimes the researcher manipulates one variable to analyze its effect on another.

For example, this model from economics states that salary is a function of a number of factors, namely education, experience, training. This model is simple, because it doesn't present the relationships between these factors; but just states that wage depends on some combination of them.

$$\text{wage} = \text{f(education, experience, training)}$$

Another canonical example comes from political science: Riker and Ordeshook's "A Theory of the Calculus of Voting" (1968). This model demonstrates the utility, or benefit, to an individual from exerting effort to go out and vote:

$$\mathrm{U(voting)} = \mathrm{p} \cdot \mathrm{B} - \mathrm{c}$$

p = probability that a single vote will be decisive
B = net benefit from your candidate winning
c = net cost of voting

such that the utility of voting equals the probability a single vote will be decisive (p) multiplied by the benefit (B), minus the cost (c).

This mathematical model succinctly demonstrates the "paradox of voting" — the probability that an individual's vote will be decisive, p, is very small, so even for large values of the benefit (B) there is no instrumental benefit to voting.

3.4.3 Empirical or Econometric Models

Empirical models are mathematical models, but are designed to be used with data. By using statistics and careful collection of data for each variable, an empirical model can provide specific estimates of the model's values by using econometrics.

Going back to the example of wages, this is an example of an econometric model called a regression analysis (you will learn more about regression in the first year of the EMPA; you don't need to worry about regression right now).

$$\mathrm{wage} = \beta_0 + \beta_1\mathrm{educ} + \beta_2\mathrm{experience} + \beta_3\mathrm{experience} + \mathrm{u}$$

Module 4: Descriptive Statistics and Visualizing Data

Public policy analysis, as well as a large body of research in the social sciences, relies heavily on *statistics.* This field is concerned with the collection, presentation, and analysis of data. Statistics can refer to both the field of study (e.g. "this paper uses statistical analysis") and the data themselves (e.g. "this paper presents employment statistics").

When collecting data, we must be concerned with both the population we are interested in studying, and the sample of data we manage to collect. The *population* is the entire group of relevant individuals, also called the universe — for example, it could be the complete number of students in the school, or all the citizens in the country, or all the legislators in the parliament.

However, it is difficult or sometimes impossible to collect data on such a large group, so instead we examine a small part of the population, called the *sample*. A sample could be one class in a school, or a selection of 1,000 or 10,000 individuals in a country. If the sample is representative of the larger population, then we can draw conclusions about the population from the sample (and determining whether this is the case forms a large part of statistics).

This module reviews common types of data, the definitions of distributions and descriptive statistics, and more generally how to visualize data. Here, we will look at a sample of a hundred voters (N=100) in a small town, and examine descriptive statistics regarding their their age (continuous variable) and their gender (discrete, indicator variable).

4.1 Data Measurement

Recall that a variable is an entity, represented by a symbol (such as X, Y, β), that can be measured quantitatively. There are two broad types of variables: categorical, and numeric.

A *categorical variable* measures a quality or characteristic, and typically represents qualitative information. It can be mutually exclusive (no overlap between categories) or exhaustive (including all possible categories). Categories can represent a value that can be ordered or ranked, or values that cannot be organized in any particular way. For example, gender or city are both categories that have no intrinsic ordering.

A *numeric variable* has values that measure quantitative information, in numbers. There are two types of numeric variables: discrete, and continuous.

> A *discrete* variable is numeric, and can only take on a certain and finite number of values. For example, the number of children a family has, or the course evaluation scores on a 5-point scale are both discrete measures. You can't have half a child, and a 5-point scale is restricted to five unique responses = {1, 2, 3, 4, 5}. Discrete variables are a function of counting.
>
> A *continuous* variable is also numeric, but one that can theoretically take any value within a given range (and can take on an infinite number of possibilities, in smaller and smaller units). For example, the height or age of an individual would be considered a continuous variable. One easy way to remember continuous variables is that they are involved in measurement.

Note: One type of categorical variable that is often treated as numeric and discrete in statistics is a "dummy" variable, also called an indicator variable. This takes on the values of 0 or 1, for two mutually exclusive categories — for example, 0=Male and 1= Female, or 0=Democrat and 1=any other party. We will discuss this further in the math pre-sessional.

Within these main types of variables, there are also many levels of measurement:

Nominal (categorical): a set of mutually exclusive and exhaustive categories, where the numerical values just "name" each category without providing additional information.

- For example, party affiliation is a nominal variable — you could assign the value of 1 to Republican, 2 to Democrat, and 3 to independent, but one number is not better than the other. This is by far the "weakest" level of measurement.

Ordinal (categorical): a set of mutually exclusive categories that can be rank ordered.

- For example, level of education completed is often an ordinal variable — 0 is "less than high school," 1 is "high school," 2 is "college," and 3 is "postgraduate." However, the distance between the numbers has no significance (the distance from 0 to 1 is not necessarily the same as the distance from 2 to 3).

Interval (numeric): similar to ordinal, but the distance between the data are meaningful, and possesses equal intervals.

- For example, survey respondents are often asked to rate their feelings on various issues of public policy, on a scale from 1 to 5 (1 being the least supportive, and 5 being the most supportive). Here, the values have equal intervals (the distance between 1 and 2, and between 4 and 5, is the same).

Ratio (numeric): has all the characteristics of the prior levels, and the distance between the data are meaningful, but includes an absolute zero (where none of the quantity being measured exists).

- For example, income, years of education, or country GDP are all ratio variables.

Variable measurement is effectively a hierarchy. Measures at the interval or ratio level are the most desirable, because they allow us to use descriptive statistics such as means and standard deviations. We will return to these at the end of the module.

4.2 Distributions

The *distribution* of a dataset presents information about all the values in the data and how frequently they occur. We care both about the center of a distribution (the mean, or average value), and how spread out the rest of the data are around the center.

One simple way to summarize a set of data is to divide the data into categories, and determine the frequency, or number of individuals belonging to each category. The arrangement of data by category with the corresponding frequencies is called the *frequency distribution*.

Age	Frequency	Percent
18-25	17	17
26-35	13	13
36-45	20	20
46-55	28	28
56-65	13	13
65 and over	9	9

This could be represented by a frequency table. Using our example, we can divide the ages of our voters into a series of categories to see the distribution. Since our sample is 100 voters, the frequency and percent are identical.

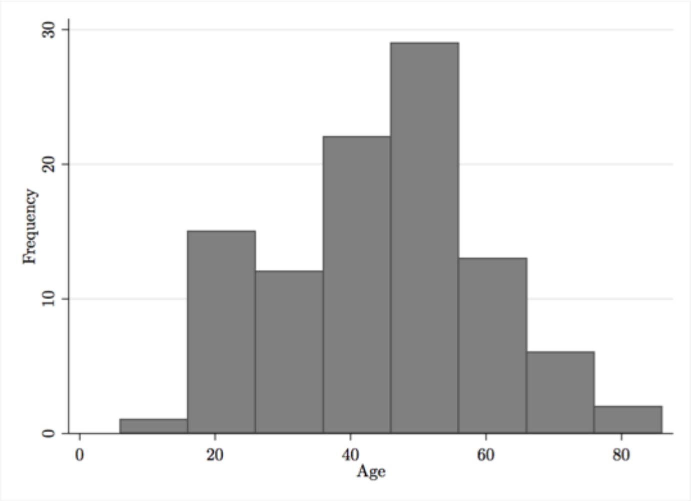
Here we can see that the category with the largest number of voters (28 voters, or 28% of our sample) is in the 46-55 age bracket. Our smallest category of voters is ages 65 and older; there are only 9 voters (or 9% of the sample) in the data.

However, it's often much easier to graphically present the distribution of data.

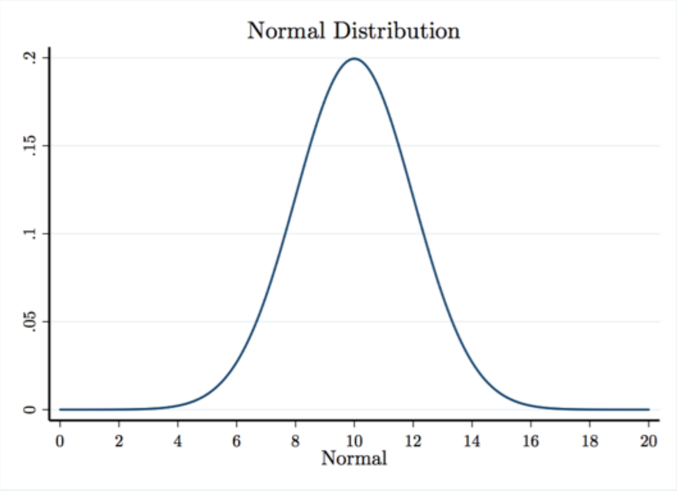
A *histogram* is a graphic representation of a frequency distribution, for a single variable. The x-axis represents the values of age, and the y-axis measures its frequency, or the number of times this value was observed in the data. The rectangle bars, sometimes called bins, can be adjusted to various categories.

For example, say we wanted to look at the distribution of voter ages in 5-year categories instead of roughly ten year categories (as in the frequency table above). This allows us to see both where the data is centered (here, approximately 43), and how much variation there is in the data across categories.

This histogram allows us to quickly see the where the greatest number of respondents are, and how much variation there is in the data across categories.



Often you will see distributions presented using a curve:



One of the most well-known distributions is called the *normal distribution,* which has data which is symmetrical around the mean. This is also known as a bell-shaped curve.

4.3 Descriptive Statistics

Descriptive statistics provide information about the basic features of the data for any sample. They are an important precursor to more advanced quantitative or statistical analysis.

These can be divided into two major categories:

1) Measures of Central Tendency

2) Measures of Dispersion

4.3.1 Measures of Central Tendency

The measures of central tendency describe a distribution in terms of its most typical, or representative, data value; these are the *mean*, the *median*, and the *mode*.

Mean

The *mean* of a set of data is calculated by taking the sum of the data, and then dividing the sum by the total number of values in the set. A mean is commonly referred to as an average, or denoted by X with a bar above it (called x-bar).

For example, take the respective ages of group of 5 voters — 25, 31, 43, 55, and 71. Sum the ages, and divide by the number of individuals in the group, to get an average age of 45.

$$\overline{X} = \frac{25 + 31 + 43 + 55 + 71}{5}$$

Median

The *median* is the middle value of an odd set of numbers arranged in order of magnitude. In the case of an even set, it is the mean of the two middle values.

In our 5 voter group example, the set of ages arranged in order is 25, 31, 43, 55, and 71. The median, or middle value, is 43. Note the median and mean are not necessarily the same —here the median is 43, while the mean is 45.

If we were to add another voter to the group, aged 80, then the median would no longer be 43. Our group would be 25, 31, 43, 55, 71, and 80; and the median would be the average of 43 and 55, so 49.

Mode

The *mode* of a set of numbers is the value that occurs most often (with the greatest frequency). If no number occurs with any frequency more than one, it has no mode.

For example:

1. 5, 6, 6, 7, 8, 9 has mode 6, and is unimodal.
2. 5, 6, 6, 7, 7, 8, 9 has two modes, 6 and 7, and is bimodal.
3. 25, 31, 43, 55, 71 has no mode.

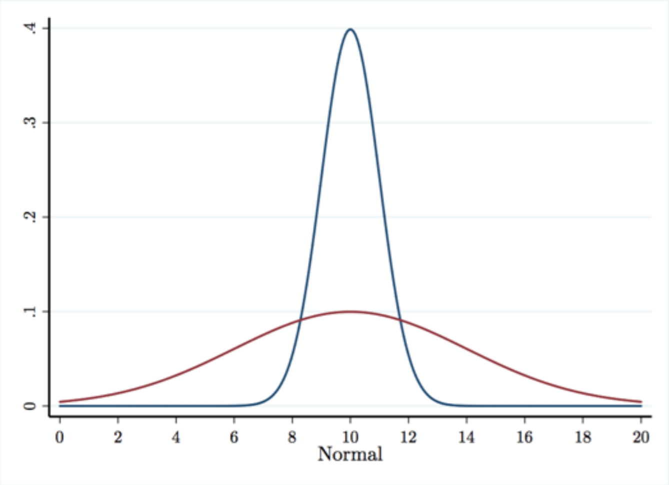
4.3.2 Dispersion

Dispersion describes how spread out a set of data is. It is important to see how the data are dispersed around the measures of central tendency.

The higher the measures of dispersion, or the more the values are spread or scattered about, the more variation there is in the data. Low dispersion means that most of the values are close to the average.

Measures of dispersion often tell us the true shape of the distribution. This is particularly important because distributions can have identical measures of central tendency, but different dispersion patterns.

This figure below shows two distributions, one red and one blue, with identical means (at 10) but varying degrees of dispersion.



The blue distribution has most of its values clustered around the average and has low dispersion; in contrast, the values representing the red distribution are scattered across all the possible values on the x axis.

The most common measures of dispersion are the range, variance, and standard deviation.

Range

The *range* is the simple difference between the lowest and highest value in a dataset. For example, for our age variable, the oldest voter in our data is 81 and the youngest is 18. The range is the highest value minus the lowest, so 81-18=63. The higher the range, the more variation in voter age in our data.

However, the range is a relatively simplistic measure of dispersion, since it is based only on the two most extreme values (also called outliers) in the dataset. For example, imagine that all voters in the data are under the age of 50 except for one individual who was 81 — in this, the range of 63 may not be typical of the dataset because of one outlier data point. To alleviate this problem, often one takes an inter-quartile range, or relies on other measures of dispersion.

Standard Deviation and Variance

The *standard deviation* measures the “average deviation,” or the amount every value in the dataset differs from the mean value.

Effectively, this tells us how spread out the values are from the mean. If the data are spread out far from the mean, the standard deviation will be large. If the data are bunched tightly together around the mean, the standard deviation will be small.

The standard deviation is typical denoted by the Greek symbol sigma (σ) or the abbreviation SD, and is calculated by taking the square root of the variance.

The *variance* is the average of the squared differences from the mean and is denoted by sigma squared. EC455 will review how to calculate this measure in depth, and no need to memorize this, but the formula is here:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

n is the number of data points

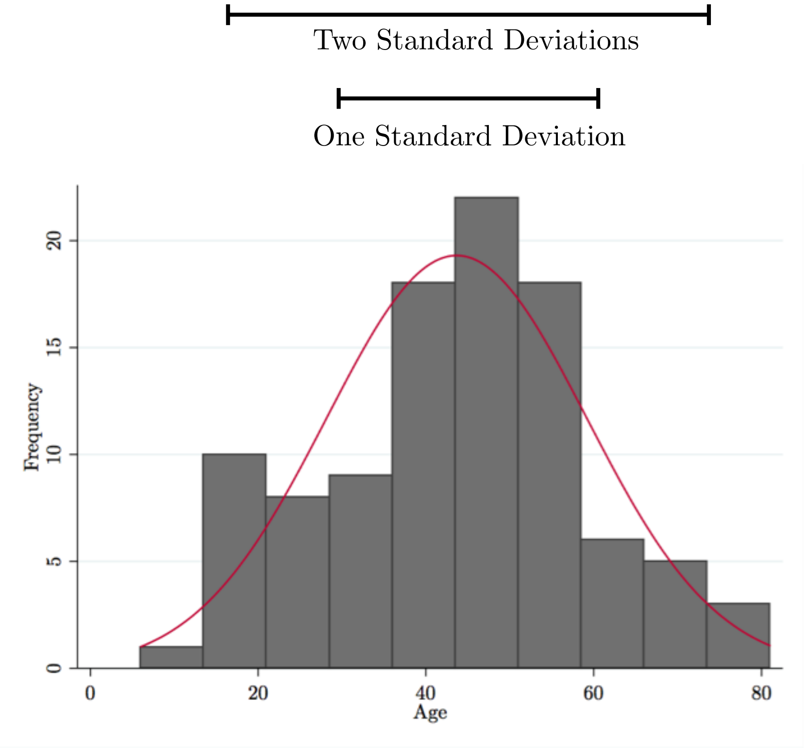
x_i represents each value of the data

$\bar{x}$ is the mean of all the values of x_i

For datasets that have a normal distribution, the standard deviation can be used to determine the proportion of values that lie within a particular range of the mean value. For such distributions it is always the case that 68% of values are less than one standard deviation (1SD) away from the mean value, that 95% of values are less than two standard deviations (2SD) away from the mean, and that 99% of values are less than three standard deviations (3SD) away from the mean.

In our age example, the mean is 43.7 and one standard deviation is 15.5.
As a result, we know that:

- 68% of the age values in the dataset will lie between MEAN - 1 SD (43.7 - 15.5 = 28.2) and MEAN + 1 SD (43.7 + 15.5 = 59.2)
- 95% of the age values in the dataset will lie between MEAN - 2 SD (43.7 - 31 = 12.7) and MEAN + 2 SD (43.7 + 31 = 74.7)



Further, in a different example, if the mean remains 43.7 but the standard deviation decreased — for example, if one standard deviation was 10 instead of 15.5 — this means the values are much less dispersed and all are closer to the mean value.

4.4 Descriptive Statistics and Levels of Measurement

Recalling our levels of measurement, some variables lend themselves to certain types of descriptive statistics.

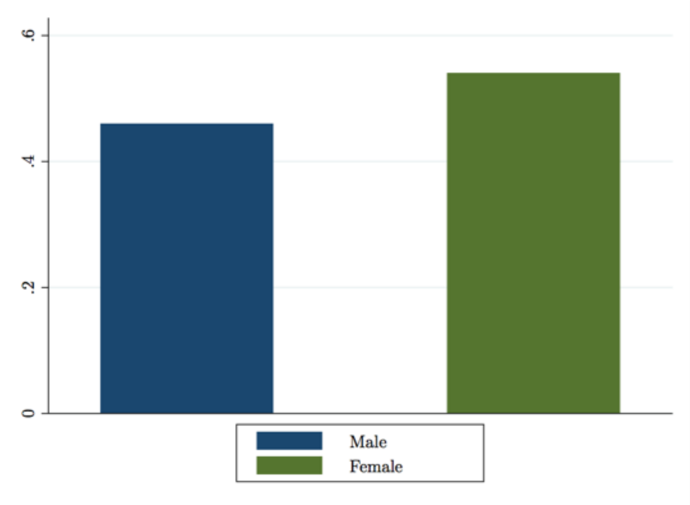
For example, it doesn't make sense to take the mean of a nominal (or categorial) variable — if voters in the sample are Republicans (=1), Democrats (=2), or Independents (=3), the average of the numbers 1, 2, and 3 don't tell us anything about party affiliations. The mode, or the frequency of party affiliations, would be much more informative.

In our district example, the average voter age is a ratio variable, however, and the mean voter age would be meaningful.

The table below presents the levels of measurement and measures of central tendency that are appropriate for each.

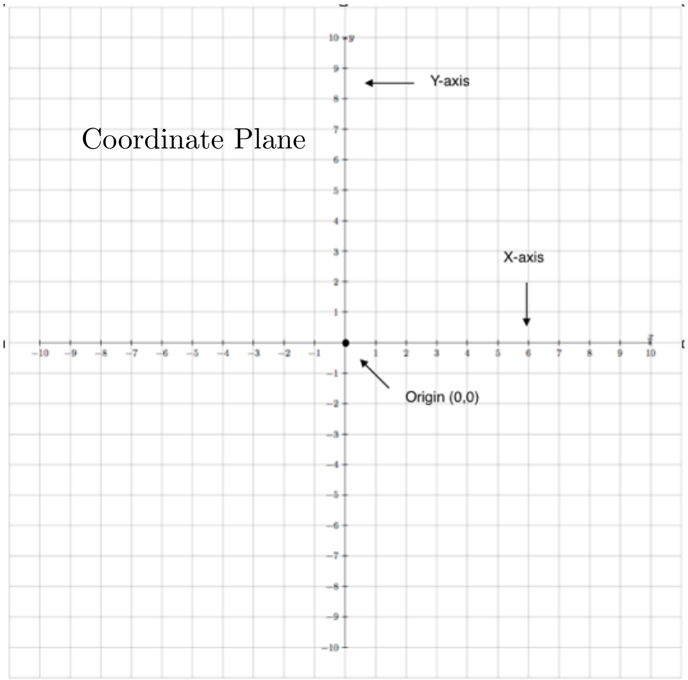
	Mode	Median	Mean	Variance/SD
Nominal (categories)	yes	no	no	no
Ordinal (ranked categories)	yes	yes	no	no
Interval/Ratio (meaningful distance and measurement)	yes	yes	yes	yes

Further, indicator variables (dummy variables that only take a value of 0 or 1) have a useful property when analyzing the mean. This bar chart presents data on the gender of the voters in our example. Here, the x-axis represents the two categories, either male or female. The y axis represents the average value for each category.



However, since this is an indicator variable, which can only take the values of 0 or 1, the average tells us the proportion of the category in the sample. Here we can see that the average for female is .54, so 54% of our sample is female. Here, we are treating the nominal indicator variable as an interval level variable.

4.5 Graphing Equations and Functions



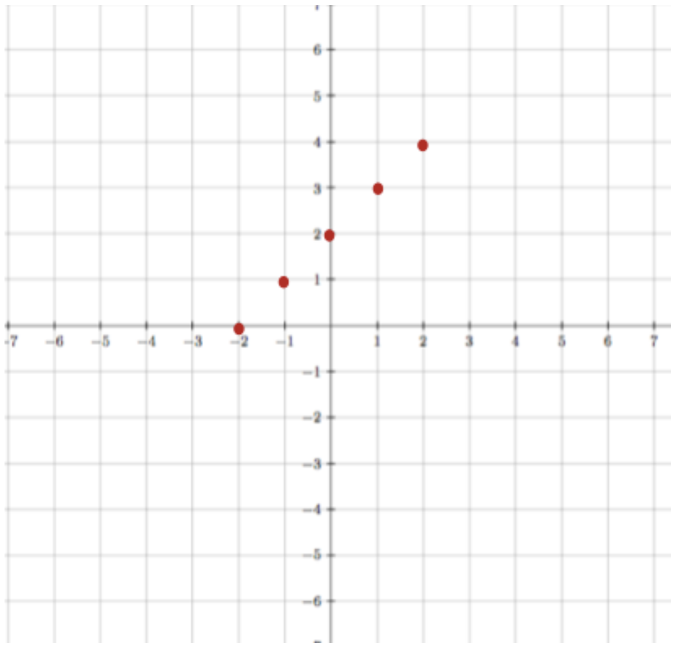
We graph equations on a *coordinate plane.* It consists of a horizontal number line, called the x axis, and a vertical number line, called the y axis. These two axes intersect at a point called the origin.

The graph of an equation is a connected set of points that all are solutions to the equation. We call each point an ordered pair of numbers (x.y), also called a coordinate.

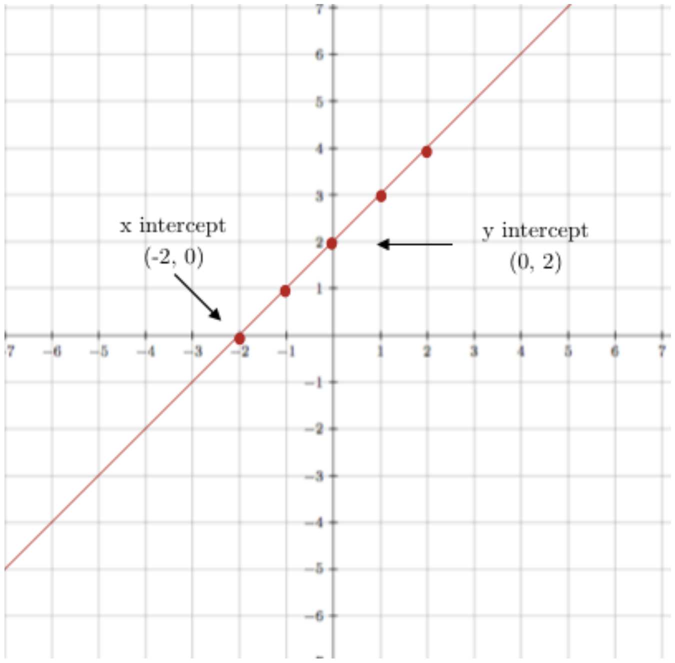
To graph an equation or function, plot a sequence of points such that a pattern emerges, and connect the points. Begin by plugging in assorted values for x, then solving the equation for y to get a set of ordered pairs.

For example: graph y = x + 2.

x	y=x + 2	Ordered pair (x,y)
-2	-2 + 2 = 0	(-2, 0)
-1	-1 + 2 = 1	(-1, 1)
0	0 + 2 = 2	(0, 2)
1	1 + 2 = 3	(1, 3)
2	2 + 2 = 4	(2, 4)



Here we have simply plotted the points from the table above.



Here we see the points connected by a straight line; this represents the function.

The point in which the graph crosses the x-axis is called the x-intercept and the point in which the graph crosses the y-axis is called the y-intercept. The x-intercept is found by finding the value of x when y = 0, at the coordinate (x, 0), and the y-intercept is found by finding the value of y when x = 0, at the coordinate (0, y).

4.5.1 Visualizing Different Types of Functions

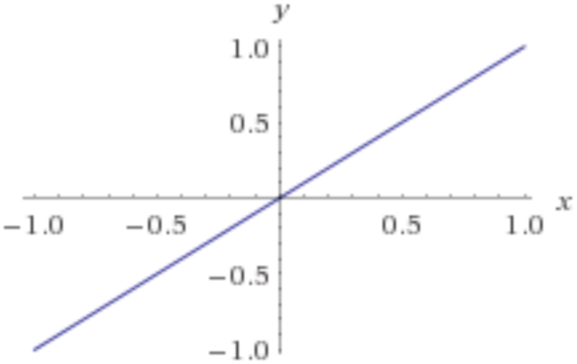
This section presents functions you might encounter in your EMPA coursework. There is no need to memorize these figures; they are just for reference.

The website https://www.wolframalpha.com is also great online resource to practice visualizing functions.

Linear functions

y = mx + b

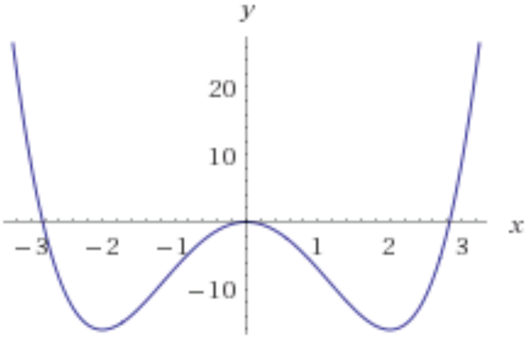
y=x *(shown)*



Polynomial functions

$y = a_n x^n + a_{n-1} x^{n-1} + \ ... \ + a_2 x^2 + a_1 x + a_0$

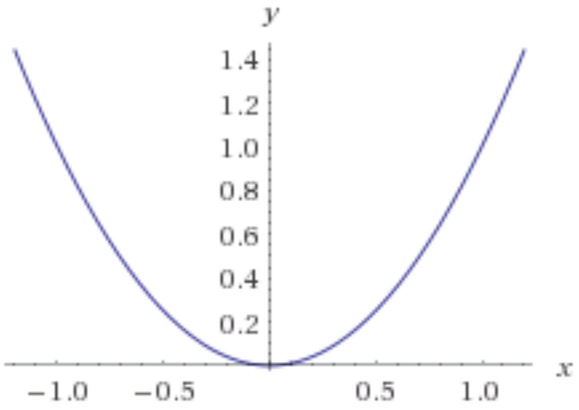
y=x^4 - 8x^2 *(shown)*



Quadratic functions

$y = ax^2 + bx + c$

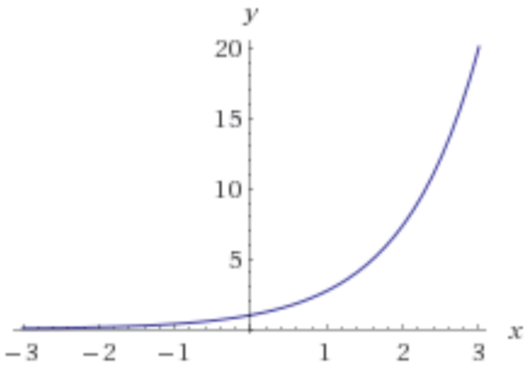
$y = x^2$ *(shown)*



Exponential functions

$y = ab^x$

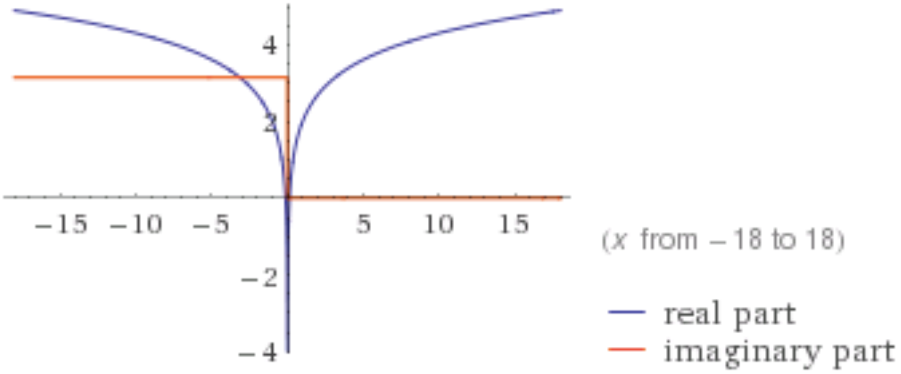
$y = e^x$ *(shown)*



Logarithmic functions

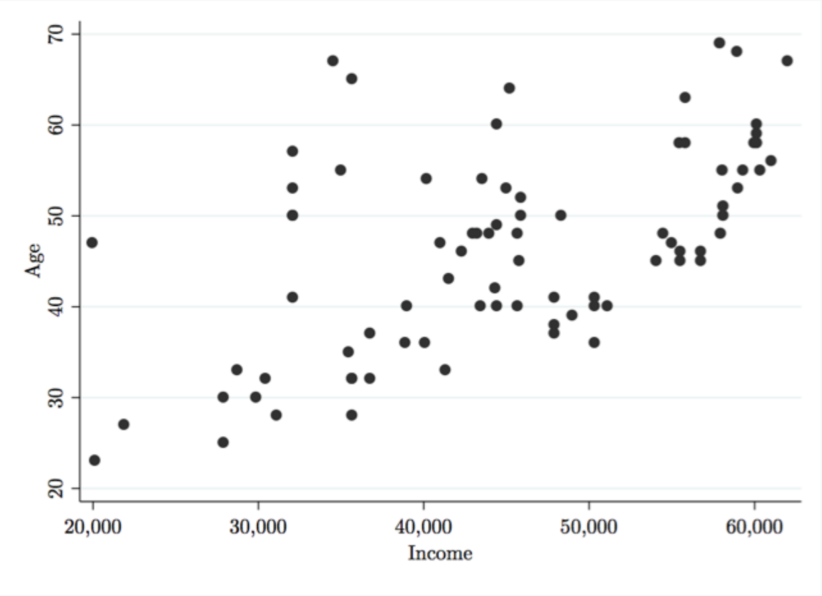
$y = \ln(x) + b$

$y = \ln(x) + 2$ *(shown)*



4.5.2 Scatterplots

Similarly, a *scatterplot* demonstrates the relationship between two variables. A scatterplot is based on the coordinate plane, by plotting ordered pairs where the values of the first variable are represented by the x-axis and the values of the second variable are represented by the y-xis.



Here, this figure plots the age (on the y-axis) and yearly income (on the x-axis) of each employed individual of working age in our dataset.

While there is a good deal of variation in income, we can see that generally there is a positive relationship — as age increases, yearly income increases. Further, scatterplots are useful for seeing if there are any outliers in the data, or any combinations that are substantively unexpected. For example, while it is possible that a 48 year old could be making $20,000 a year, we might want to double-check this observation just to be sure.

4.5.3 Visualizing Correlations

As mentioned earlier, the correlation between two variables measures to what extent the variables move in the same way (a positive correlation) or move in opposite ways (negative correlation). This relationship can be plotted as a scatterplot, or expressed as a correlation coefficient. The correlation coefficient is between -1 and 1, with -1 being a perfect and strong negative correlation and similarly 1 being a perfect and strong positive correlation.

Recall that a correlation is positive when the values of two variables increase together; visually, this looks like a line with an increasing slope (or gradient). A correlation is negative when one value decreases as the other increases; this looks like a line with a negative slope. No correlation just resembles a handful of seemingly unrelated points.

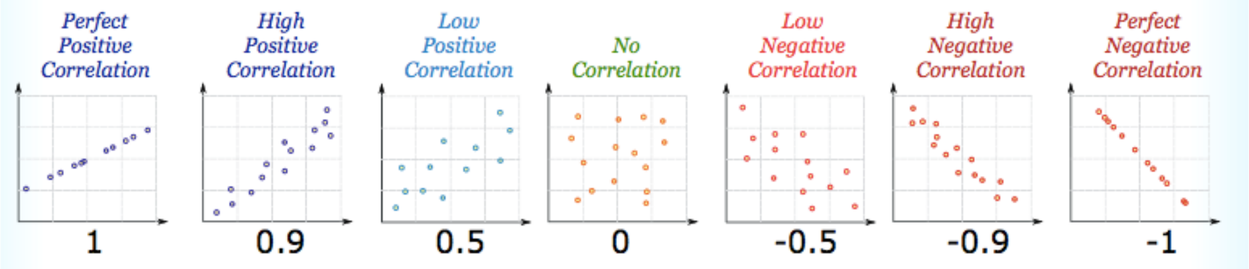
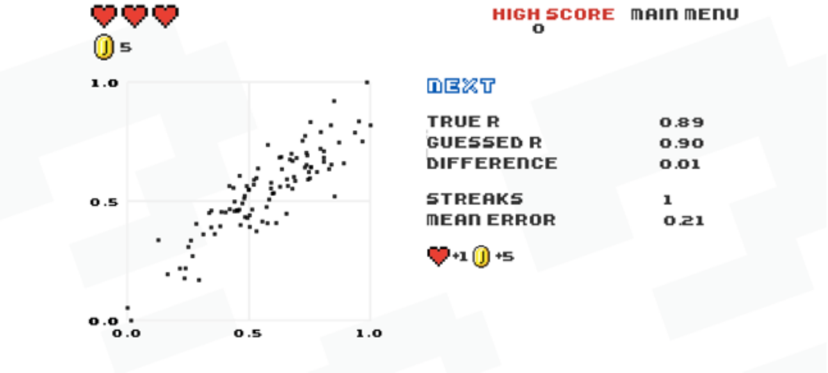


Image courtesy of http://www.mathsisfun.com/data/correlation.html.



TRUE R 0.89
GUESSED R 0.90
DIFFERENCE 0.01

STREAKS 1
MEAN ERROR 0.21

For more on visualizing correlations, this website provides an entertaining way to practice, in the form of a nostalgic 8-bit video game:

http://guessthecorrelation.com