1 Preliminaries: Statistical and Causal Models

Statistics generally concerns itself not with absolutes but with **likelihoods**, thus the language of probability is extremely important to it.

Probability is similarly important to the study of causation because most causal statements are uncertain

“**careless driving causes accidents**”

which is true, but does not mean that a careless driver is certain to get into an accident.

Probability is the way we express uncertainty, even if many other approaches are available to manage it.

In this course, we will use the language and laws of probability to express our **belief and uncertainty about the world**.

We provide a glossary of the most important terms and concepts they will need to know in order to understand the rest of the course.

A **variable** is any property or descriptor that can take multiple values.

Probability of multiple values at once.

$$P(Age = 38, Gender = male)$$

A study to **compare health of smokers and nonsmokers**


- *Age*
- *Gender*
- *Family history of cancer?*
- *How many years smoking?*

An individual randomly selected from the population is aged 38.

A **Variable** can be thought of as a **Question**, to which the **Value** is the **Answer**.

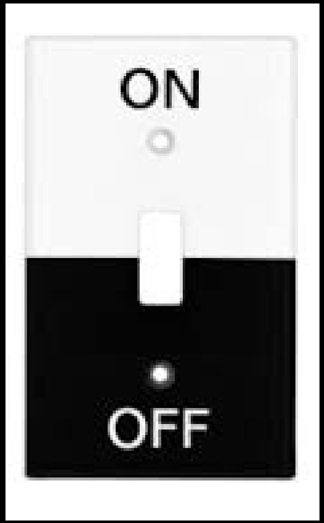
$$P(Age = 38)$$

Question:	*How old is the participant?*	**Variable:**	Age	X
Answer:	*38 years old*	**Value:**	38	x

$$P(X = x) \qquad P(x)$$

A **variable** can be

- **discrete** or **categorical**; can take one of finite or countably infinite set of values in any range.



Light switch

- **continuous**; can take any one of an infinite set of values on a continuous scale.

Person's weight

An **event** is any assignment of a value or set of values to a variable or set of variables.

Examples of event

- $X = 1$
- $X = 1$ OR $X = 2$
- $X = 1$ AND $Y = 3$
- $X = 1$ OR $Y = 3$

The patient recovers

coin flips lands on heads

Variable: the patient's status

Value: recovered

Variable: coin flips

Value: head

Another way of thinking of an **event** in probability is this:

Any declarative statement (a statement that can be true or false) is an event.

The probability that some **event *A*** occurs, given that we know some other **event *B*** has occurred, is the **conditional probability of *A* given *B***.

X

$P(X = x|Y = y)$ $P(x|y)$ $Flu = \{yes, no\}$

Y

$Temperature\ C° = 39$

The probability we assign to the **event *"X = x"*** changes drastically, depending on the **knowledge *"Y = y"*** we condition on.

$P(yes) = 0.01$ $P(yes|39) = 0.65$

When dealing with probabilities represented by frequencies in a data set, one way to think of conditioning is filtering a data set based on the value of one or more variables.

In **Table 1.3**, there were 132,949,000 votes cast in total, so we would estimate that the probability that a given voter was younger than the age of 45 is

$$P(Voter's\ Age < 45) = \frac{20{,}539{,}000 + 30{,}756{,}000}{132{,}949{,}000} = 0.3858$$

Age of U.S. voters in the 2012 presidential election.

TABLE 1.3 Age breakdown of voters in 2012 election (all numbers in thousands)

Age Group	# of voters
18-29	20,539
30-44	30,756
45-64	52,013
65+	29,641
	132,949

1.3.3 Probability and Statistics: Conditional Probability

When dealing with probabilities represented by frequencies in a data set, one way to think of conditioning is filtering a data set based on the value of one or more variables.

In **Table 1.3**, there were 132,949,000 votes cast in total, so we would estimate that the probability that a given voter was younger than the age of 45 is

$$P(Voter's\ Age < 45) = \frac{20{,}539{,}000 + 30{,}756{,}000}{132{,}949{,}000} = 0.3858$$

Age of U.S. voters in the 2012 presidential election.

TABLE 1.3 Age breakdown of voters in 2012 election (all numbers in thousands)

Age Group	# of voters
18-29	20,539
30-44	30,756
	132,949

Suppose, however, we want to estimate the probability that a voter was younger than the age of 45, **given that we know** he was elder than the age of 29.

To find this out, we simply filter the data to form a new set (**Table 1.4**), using only the cases where the voters were older than 29.

In this new data set, there are 112,410,000 total votes, so we would estimate that

$$P(Voter's\ Age < 45 | Voter's\ Age > 29) = \frac{30{,}756{,}000}{112{,}410{,}000} = 0.2736$$

TABLE 1.4 Age breakdown of voters over the Age of 29 in 2012 election (all numbers in thousands)

Age Group	# of voters
30-44	30,756
	112,410

TABLE 1.3 Age breakdown of voters in 2012 election (all numbers in thousands)

Age Group	# of voters
18-29	20,539
30-44	30,756
45-64	52,013
65+	29,641
	132,949

Conditional probabilities such as these play an important role in **investigating causal questions**, as we often want to compare how the probability (or equivalently, risk) of an outcome changes under different filtering, or exposure, conditions.

How does the probability of developing lung cancer for smokers compare to the analogous probability for nonsmokers?

It might happen that the probability of one event remains unaltered with the observation of another.

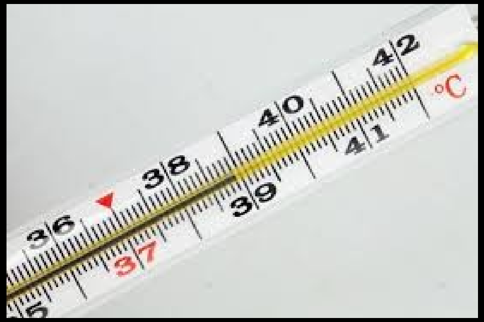
X

$Flu = \{yes, no\}$

$P(yes) = 0.01$

Y

$Temperature\ C° = 39$



$P(yes|39) = 0.65$

Z

$Age = 27$

$P(yes|27) = 0.01$

Two **events *A* and *B* are said to be independent** if

$$P(A|B) = P(A)$$

The knowledge that *B* has occurred gives us no additional information about the probability of ***A*** occurring.

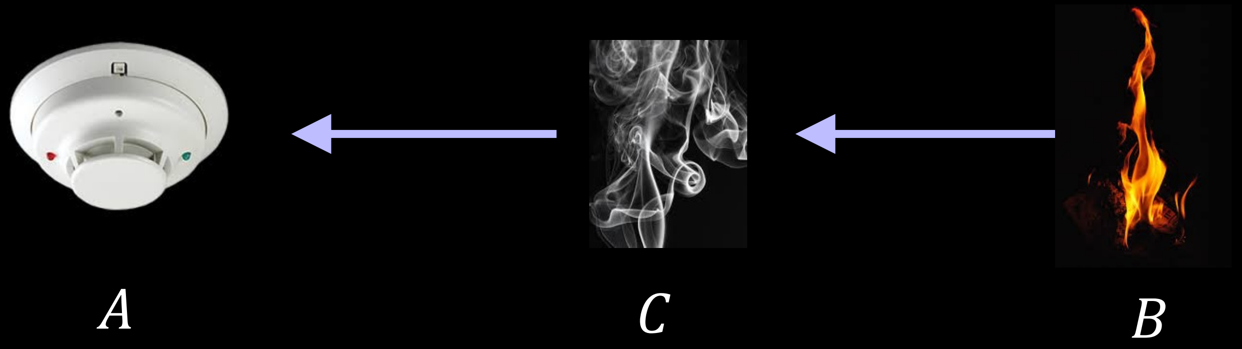
If the following equality

$$P(A|B) = P(A)$$

does not hold, then ***A* and *B* are said to be dependent**.

Two events ***A* and *B*** are **conditionally independent given** a third **event *C*** if

$$P(A|B,C) = P(A|C) \qquad P(B|A,C) = P(B|C)$$



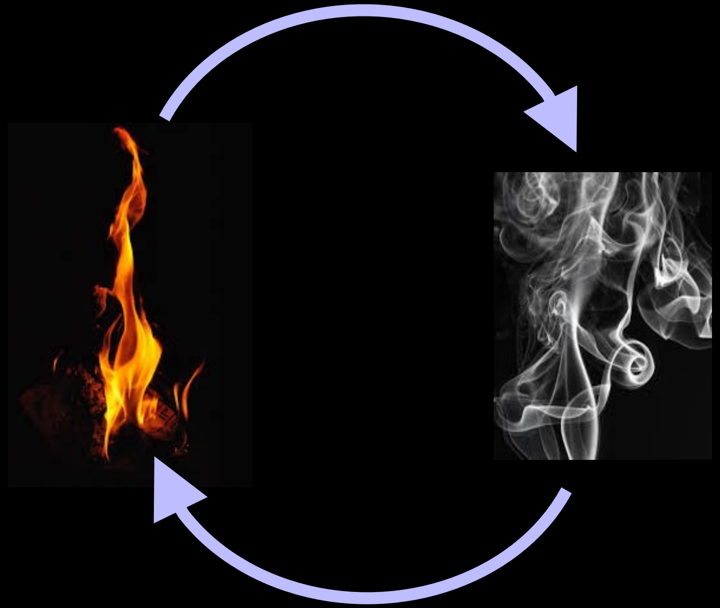
Dependence and **independence** are **symmetric relations**

- If ***A* is dependent on *B***, then ***B* is dependent on *A*.**

$$P(A|B) \neq P(A) \Leftrightarrow P(B|A) \neq P(B)$$

- If ***A* is independent on *B***, then ***B* is independent on *A***.

$$P(A|B) = P(A) \Leftrightarrow P(B|A) = P(B)$$



Variables, like events, can be **dependent** or **independent** of each other.

Two **variables *X* and *Y* are considered independent** if for every value *x* and *y* that *X* and *Y* can take, we have

Independence of variables **is a symmetrical relation**

$$P(X=x|Y=y)=P(X=x) \qquad X \perp Y$$

$$P(Y=y|X=x)=P(Y=y) \qquad Y \perp X$$

$$P(X=x,Y=y)=P(X=x)\,P(Y=y)$$

If for any pair of values of *X* and ***Y***, one of the above equalities does not hold, then ***X* and *Y* are said to be dependent**.

$$X \not\perp Y \qquad Y \not\perp X$$

Independence of variables can be understood as a set of independencies of events.

Variables, like events, can be **conditionally dependent** or **conditionally independent** of each other given some other variables.

Two **variables *X* and *Y* are considered conditionally independent, given a third variable *Z***, if for every value *x* and *y* that *X* and *Y* can take, for each value *z* that *Z* can take

$$P(X = x|Y = y, Z = z) = P(X = x|Z = z) \qquad X \perp Y|Z$$

$$P(Y = y|X = x, Z = z) = P(Y = y|Z = z) \qquad Y \perp X|Z$$

$$P(X = x, Y = y|Z = z) = P(X = x|Z = z)\, P(Y = y|Z = z)$$

A **probability distribution** for a variable *X* is the set of probabilities assigned to each possible value of *X*.

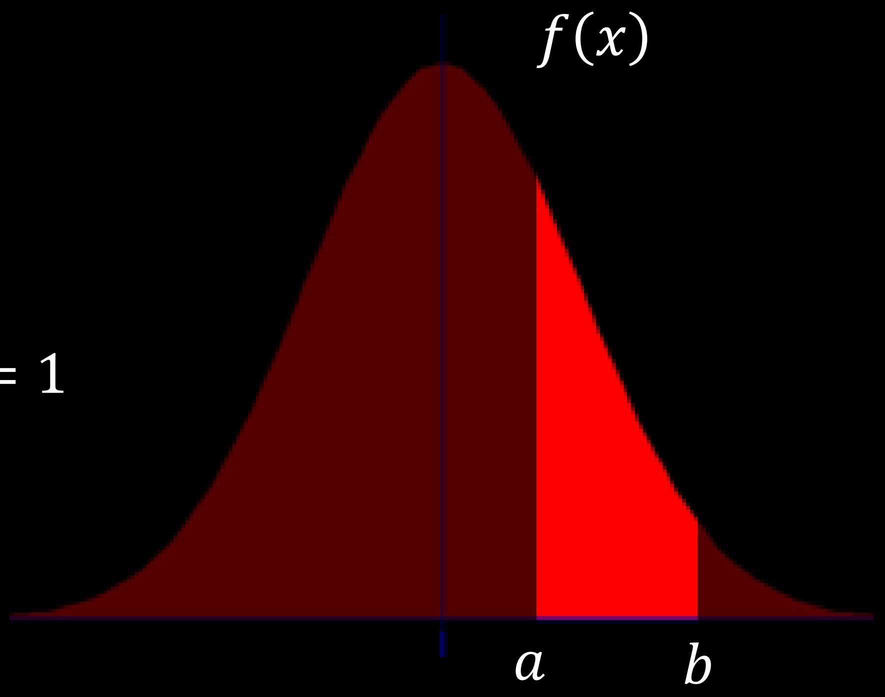
$X \in \{1, 2, 3\}$ $\quad P(X = 1) = 0.50, \quad P(X = 2) = 0.25, \quad P(X = 3) = 0.25$

$P(X = 1) = 0 \quad \Rightarrow \quad X = 1 \quad \text{is an impossible event}$

$P(X = 2) = 1 \quad \Rightarrow \quad X = 2 \quad \text{is the certain event}$

Continuous variables also have probability distributions, typically represented by a function f called **density function** and such that

$$\int_{-\infty}^{+\infty} f(x)\, dx = 1$$



***X* and *Y* Independent**

$$f(x|y) = f(x) \qquad f(y|x) = f(y) \qquad f(x, y) = f(x)\, f(y)$$

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

A **probability distribution** for a variable *X* is the set of probabilities assigned to each possible value of *X*.

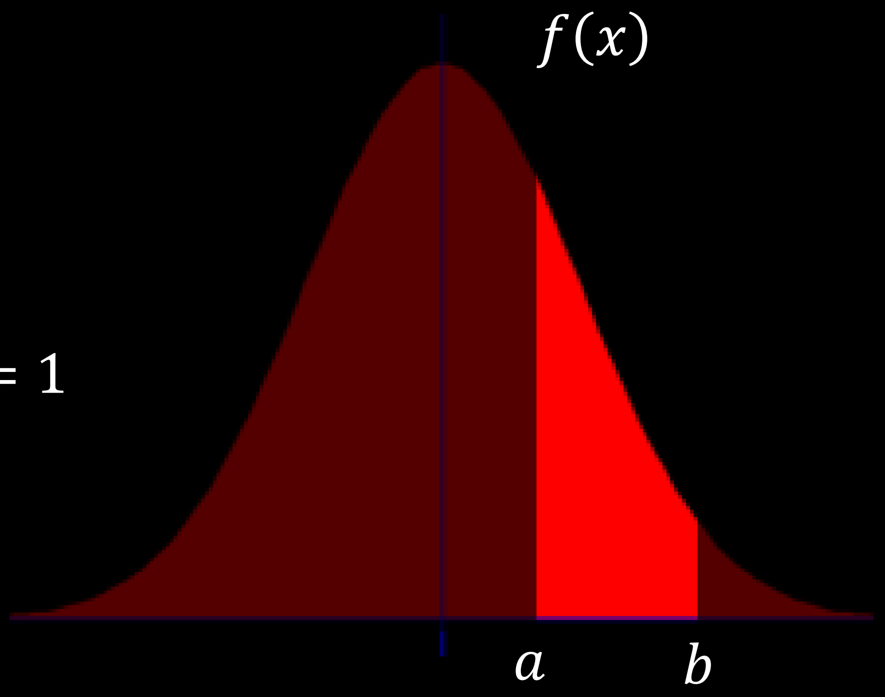
$X \in \{1, 2, 3\}$ $\quad P(X = 1) = 0.50, \quad P(X = 2) = 0.25, \quad P(X = 3) = 0.25$

$P(X = 1) = 0 \quad \Rightarrow \quad X = 1$ is an impossible event

$P(X = 2) = 1 \quad \Rightarrow \quad X = 2$ is the certain event

Continuous variables also have probability distributions, typically represented by a function f called **density function** and such that

$$\int_{-\infty}^{+\infty} f(x)\, dx = 1$$



$$P(a \le X \le b) = \int_a^b f(x)\, dx$$

Joint Probability Distribution $\quad X \in \{1,2\} \quad Y \in \{1,2\}$

$P(X = 1, Y = 1) = 0.2 \quad P(X = 1, Y = 2) = 0.3$

$P(X = 2, Y = 1) = 0.4 \quad P(X = 2, Y = 2) = 0.1$

1.3.6 Probability and Statistics: The Law of Total Probability

There are several universal probability truths that are useful to know.

Given any **pair *A* and *B* of mutually exclusive events** (i.e., ***A*** and ***B*** can not co-occur), we have

$$P(A \text{ or } B) = P(A) + P(B)$$

For any **pair *A* and *B* of events**, we have

$$P(A) = P(A \text{ and } B) + P(A \text{ and } \overline{B})$$

In general, for any set of events

$$B_1, B_2, \dots, B_n$$

such that exactly one of them must be true (it forms a partition), we have **the law of total probability**

$$P(A) = P(A, B_1) + P(A, B_2) + \cdots + P(A, B_n)$$

Furthermore, we know the following

$$P(A, B) = P(A|\, B)\, P(B)$$

$$P(A|\, B) = \frac{P(A, B)}{P(B)}$$

$$P(A|B) = P(A)$$

$$P(A, B) = P(A)\, P(B)$$

A relevant formula is the **Bayes' rule or formula**, which can be derived as follows

$$P(A,B) = P(A|\,B)\,P(B) \qquad P(B,A) = P(B|\,A)\,P(A)$$

$$P(A,B) = P(B,A) = \boxed{P(A|\,B)\,P(B) = P(B|\,A)\,P(A)}$$

$$\boxed{P(A|\,B) = \frac{P(B|\,A)\,P(A)}{P(B)}}$$

We can write a different form for **the law of total probability**

$$P(A) = P(A,B_1) + P(A,B_2) + \cdots + P(A,B_n)$$

$$P(A) = P(A|B_1)\,P(B_1) + P(A|B_2)\,P(B_2) + \cdots + P(A|B_n)\,P(B_n)$$

Useful because, often we will find ourselves in a situation where we cannot assess $P(A)$ directly, but we can through this decomposition.

Indeed, it is generally easier to assess conditional probabilities such that $P(A|B_k)$, which are tied to specific contexts, rather than $P(A)$, which is not attached to a context.

30% of disks

one out of 5,000 are defective (*D*)

factory *A*

70% of disks

one out of 10,000 are defective (*D*)

factory *B*

Which is the probability that a randomly selected disk will be defective (*D*)?

$P(D)$ = ?

$$P(D) = P(D|A)\,P(A) + P(D|B)\,P(B)$$

$$= \frac{1}{5{,}000}0.3 + \frac{1}{10{,}000}0.7$$

$$= 0.00013$$

We roll two dice, and we want to know the probability that the second roll is higher than the first

$$P(A) = P(Roll\ 2 > Roll\ 1)$$

No obvious way to calculate this probability all at once. But if we break it down into contexts

$B_1, B_2, \dots, B_6$

by conditioning on the value of the first die (B_k means the roll of the first die is *k*), it becomes easy to solve:

$$P(Roll\ 2 > Roll\ 1) = P(Roll\ 2 > Roll\ 1 | Roll\ 1 = 1)\ P(Roll\ 1 = 1) +$$

$$+ P(Roll\ 2 > Roll\ 1 | Roll\ 1 = 2)\ P(Roll\ 1 = 2) +$$

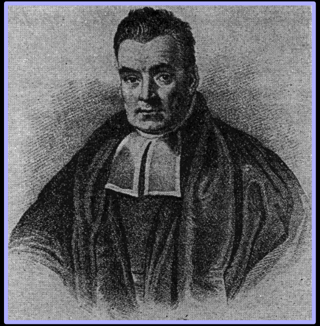
$$+ \ \dots$$

$$+ P(Roll\ 2 > Roll\ 1 | Roll\ 1 = 6)\ P(Roll\ 1 = 6)$$

$$= \left(\frac{5}{6} \times \frac{1}{6}\right) + \left(\frac{4}{6} \times \frac{1}{6}\right) + \left(\frac{3}{6} \times \frac{1}{6}\right) + \left(\frac{2}{6} \times \frac{1}{6}\right) + \left(\frac{1}{6} \times \frac{1}{6}\right) + \left(\frac{0}{6} \times \frac{1}{6}\right) = \frac{5}{12}$$

$$P(A|\, B) = \frac{P(B|\, A)\, P(A)}{P(B)}$$

When using **Bayes' rule**, we sometimes loosely refer to **event *A*** as the **hypothesis** and to **event *B*** as the **evidence**.



In many cases, we know or can easily determine $P(B|\, A)$ (probability that a piece of evidence will occur given that our hypothesis is correct)

but it's much harder to figure out $P(A|B)$ (the probability of the hypothesis being correct, given that we obtain a piece of evidence)

which is the question we most often want to answer in the real world.

probability of **evidence *B*** given that **hypothesis *A*** is correct **(likelihood)**

updated belief in **hypothesis *A*** **(posterior probability)** → $P(A|\, B) = \frac{P(B|\, A)\, P(A)}{P(B)}$

belief in **hypothesis *A*** **(prior probability)**

probability of ***B*** **(evidence)**

1.3.7 Probability and Statistics: Using Bayes' Rule

You are in a casino, and you hear a dealer shout **"11!"**.

You know that the only two games that happen to have such an outcome are:

Craps

Roulette

You know that there are as many **craps games** as **roulette games** going on at any moment, thus

$$P(craps) = P(roulette) = 0.5$$

What is the probability that the dealer is working at a game of **craps**, given that he shouted **"11!"**?

craps is the **hypothesis**

"11!" is the **evidence**

$$P(craps|"11!") = ?$$

Betting on the sum of the roll of two dice.

	⚀	⚁	⚂	⚃	⚄	⚅
⚀	2	3	4	5	6	7
⚁	3	4	5	6	7	8
⚂	4	5	6	7	8	9
⚃	5	6	7	8	9	10
⚄	6	7	8	9	10	11
⚅	7	8	9	10	11	12

craps and **roulette** are **hypothesis**

Craps

$$P(\text{"11!"}|craps) = \frac{2}{36}$$

Roulette

$$P(\text{"11!"}|roulette) = \frac{1}{38}$$

$$P(\text{"11!"}) = P(\text{"11!"}|craps)\,P(craps) + P(\text{"11!"}|roulette)\,P(roulette) =$$

$$= \frac{2}{36} \times \frac{1}{2} + \frac{1}{38} \times \frac{1}{2} = \frac{7}{171}$$

Betting on the sum of the roll of two dice.

Craps

Roulette

$$P(A|\,B) = \frac{P(B|\,A)\,P(A)}{P(B)}$$

$$P(craps|"11!") = \frac{P(11!|craps)\,P(craps)}{P("11!")} = \frac{\frac{1}{18}\times\frac{1}{2}}{\frac{7}{171}} = 0.679$$

$$P("11!") = P("11!"|craps)\,P(craps) + P("11!"|roulette)\,P(roulette) =$$

$$= \frac{2}{36}\times\frac{1}{2}+\frac{1}{38}\times\frac{1}{2} = \frac{7}{171}$$

Monty Hall Game

Behind two doors

Behind a door

Monty Hall Game

Behind two doors

Behind a door

Your Choice

You are asked to chose a door

You are offered the following alternatives

- keep your choice
- change your choice

Monty Hall Game

Your Choice

You are asked to chose a door

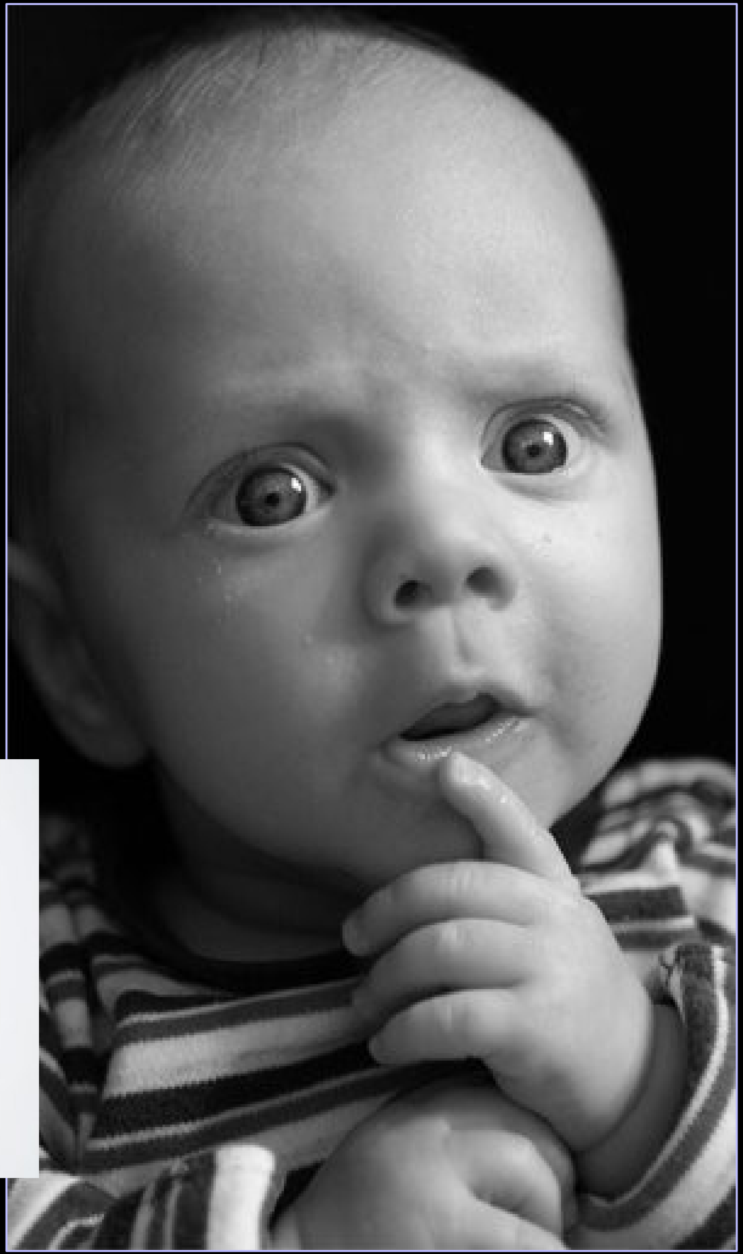
You are offered the following alternatives

- keep your choice
- change your choice

You puzzled?

What your choice?

Why?



In statistics we often deal with data sets and probability distributions that are too large to effectively examine each possible combination of values.

Instead, we use statistical measures to represent, with some loss of information, meaningful features of the distribution.

Expected Value or Mean

Can be used when the variable takes on numerical values

$$E(X) = \sum_x x\, P(X = x)$$

$$X \in \{1,2,3,4,5,6\}$$

$$E(X) = 1 \times P(1) + 2 \times P(2) + 3 \times P(3) + 4 \times P(4) + 5 \times P(5) + 6 \times P(6) = 3.5$$

Expected Value of any function of *X*, i.e. *g*(*X*)

$$E[g(X)] = \sum_x g(x)\, P(x)$$

$$g(X) = X^2 \longrightarrow E[g(X)] = 1^2 \times P(1) + 2^2 \times P(2) + 3^2 \times P(3) + 4^2 \times P(4) + 5^2 \times P(5) + 6^2 \times P(6) = 15.17$$

We can also calculate the **expected value of *Y* conditioned on *X***

$$E(Y|X=x) = \sum_y y\, P(Y=y|X=x)$$

$E(X)$ is one way to make a "best guess" of *X*'s value.

Out of all the guesses *"g"* that we can make, $g(X) = E(X)$ minimizes the expected squared error

$$E[(g(X)-X)^2] = \sum_x (g(X)-X)^2\, P(x)$$

Similarly,

$$E(Y|X=x)$$

represents a best guess of *Y*, given that we observe *X* = *x*.

If $g(Y) = E(Y|X=x)$, then the following is minimized

$$E[(g(Y)-Y)^2|X=x] = \sum_y (g(Y)-Y)^2\, P(y|x)$$

$$E(Voter's\ Age) = 23.5 \times 0.16 + 37.0 \times 0.23 + 54.5 \times 0.39 + 70.0 \times 0.22 = 48.9$$

Assumptions

- every age within each category is equally likely
- the oldest age of any voter is 75

What if we were asked to guess the age of a randomly selected voter, with the understanding that if we were off "e" years, we would lose e^2 euros?

We would lose the least money "e^2", on average, if we guessed the age to be 48.9.

Age of U.S. voters in the 2012 presidential election.

TABLE 1.3 Age breakdown of voters in 2012 election (all numbers in thousands)

Age Group			# of voters
18-29	23.5	0.16	20,539
30-44	37.0	0.23	30,756
45-64	54.5	0.39	52,013
65+	70.0	0.22	29,641
			132,949

1.3.8 Probability and Statistics: Expected Values

$$E(Voter's\ Age | Voter's\ Age < 45) = 23.5 \times 0.4 + 37.0 \times 0.6 = 31.6$$

The use of expectations as a basis for predictions or "best guesses" hinges to a great extent on an implicit assumption regarding the distribution of X or $Y|X=x$, namely that such distributions are approximately symmetric.

If, however, the distribution of interest is highly skewed, other methods of prediction may be better.

In such cases, for example, we might use the **median** of the distribution of X as our "best guess", this estimate minimizes the expected absolute error.

$$E(|g(X) - X|)$$

What if we were asked to guess the age of a randomly selected voter younger than the age of 45, with the understanding that if we were off "e" years, we would lose e^2 euros?

TABLE 1.3 Age breakdown of voters in 2012 election (all numbers in thousands)

Age Group			# of voters
18-29	23.5	0.4	20,539
30-44	37.0	0.6	30,756
			$51,295$

The **variance** of a variable *X*, denoted

$$Var(X) \quad or \quad \sigma_X^2$$

is a measure of roughly how “spread out” the values of *X* in a data set or population are from their mean.

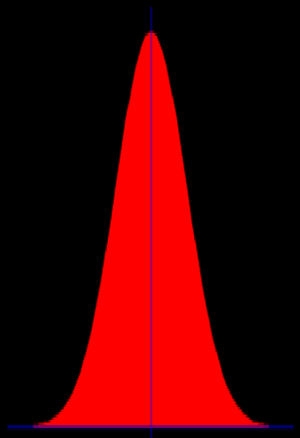
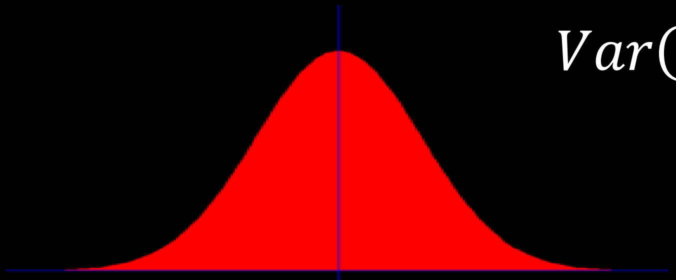


TABLE 1.3 Age breakdown of voters in 2012 election (all numbers in thousands)

Age Group	# of voters
18-29 **23.5**	**0.4** 20,539
30-44 **37.0**	**0.6** 30,756

$$Var(X) = E[(X - E(X))^2]$$
$$= E[(X - \mu)^2]$$



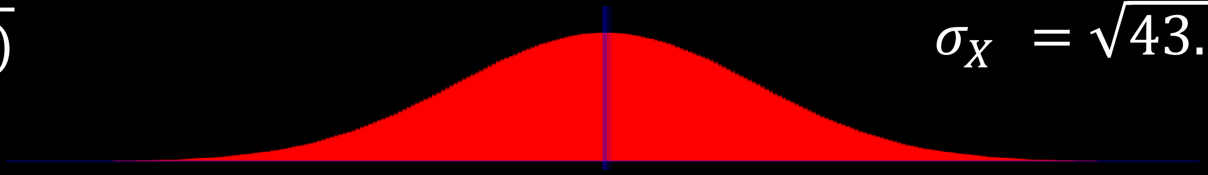
variance of under 45 voters’ age distribution

$$Var(X) = [(23.5 - 31.6)^2 \times 0.4]$$
$$+ [(37.0 - 31.6)^2 \times 0.6]$$
$$= 43.74$$

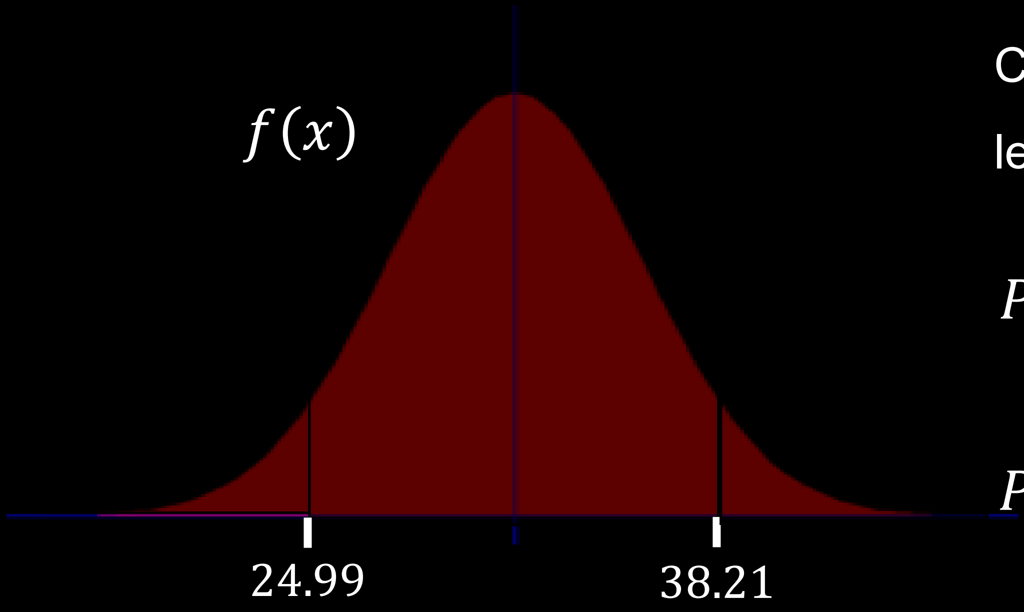
Standard Deviation

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{Var(X)}$$

$$\sigma_X = \sqrt{43.74} = 6.61 \ years$$



Expressed the same units as *X*.



$31.6 - 6.61$ $\quad$ $31.6 + 6.61$

$E(X) = 31.6 \qquad \sigma_X = 6.61 \; years$

Choosing a voter at random, chances are high that his/her age will fall less than 6.61 years away from the average 31.6.

$$P(a \le X \le b) = P(X \le b) - P(X \le a)$$

$$P(X \le a) = \int_{-\infty}^{a} f(x)\,dx \qquad P(X \le 24.99) = \int_{-\infty}^{24.99} f(x)\,dx$$

$$P(X \le b) = \int_{b}^{+\infty} f(x)\,dx \qquad P(X \le 38.21) = \int_{38.21}^{+\infty} f(x)\,dx$$

$$P(a \le X \le b) = \int_{a}^{b} f(x)\,dx \qquad P(24.99 \le X \le 38.21) = \int_{24.99}^{38.21} f(x)\,dx$$

Of special importance is the expectation of the product

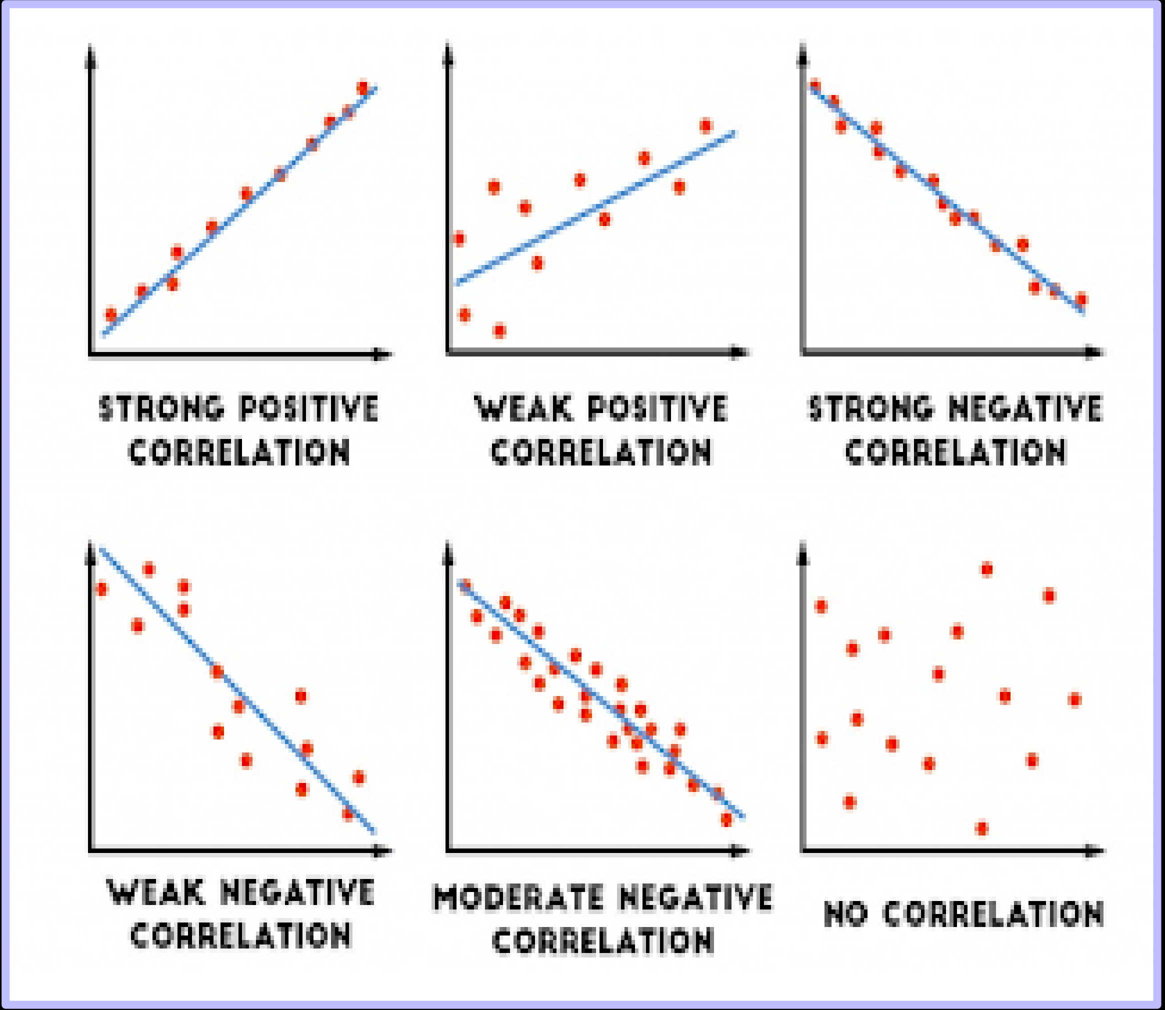
$$(X - E(X))\,(Y - E(Y))$$

which is known as the **covariance of *X* and *Y***, defined as

$$\sigma_{XY} \triangleq E[(X - E(X))\,(Y - E(Y))]$$

It measures the degree to which *X* and *Y* **covary**, that is, the degree to which the two variables **vary together**, or are **associated**.

A specific way in which *X* and *Y* covary; it **measures the extent to which *X* and *Y* linearly covary**.



Correlation between *X* and *Y*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \qquad \rho_{XY} \in [-1, +1] \qquad X \text{ and } Y \text{ independent} \Rightarrow \sigma_{XY} = \rho_{XY} = 0$$

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug			No Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Men	87	81	93%	270	234	87%
Women	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

We learned from **Simpson's Paradox** that certain decisions cannot be made on the basis of data alone, but they depend on **the story behind the data**.

We now introduce the mathematical language of **Graph Theory** where the story behind the data can be told.

Graph; consists of a collection of **nodes** (vertices) and **edges**.

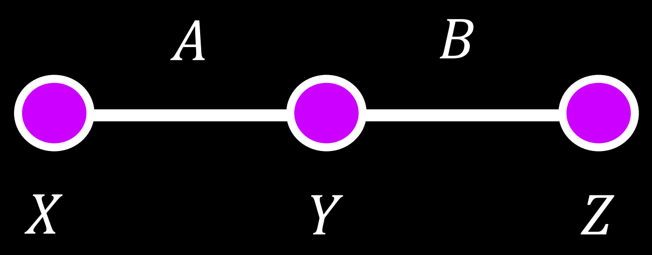
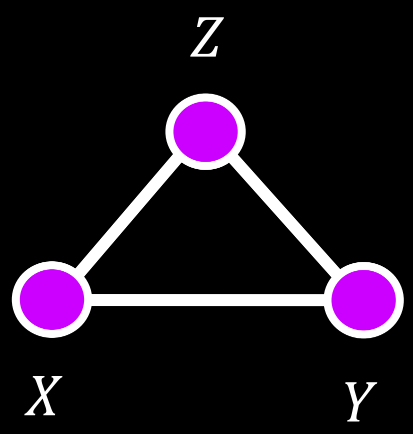


Figure 1.5

Adjacent nodes; if there is an edge between them.

Complete graph; if there is an edge between every pair of nodes.

The graph in **Figure 1.5** is not complete while it is complete the graph to the left.



X and Y are adjacent nodes, as well as Y and Z, while X and Z are not adjacent nodes.

Path between two nodes *X* and *Y*; sequence of nodes beginning with *X* and ending with *Y*, in which each node is connected to the next by an edge.

In **Figure 1.5**, there is a path between node *X* and node *Z*, because node *X* is connected to node *Y* which in turn is connected to node *Z*.

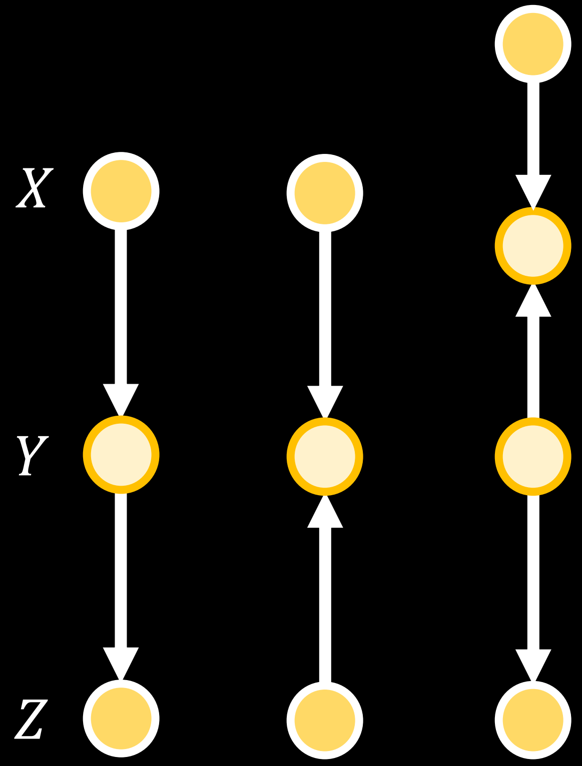
Edges can be **directed** or **un-directed**.

A graph with directed edge is a **directed graph**.

X is a **parent node** of *Y* $pa(Y) = X$

Y is a **child node** of *X* $ch(X) = Y$

A path between two nodes is a **directed path** if can be traced along the arrows, that is, if no node on the path has two edges on the path directed into it, or two edges directed out of it.



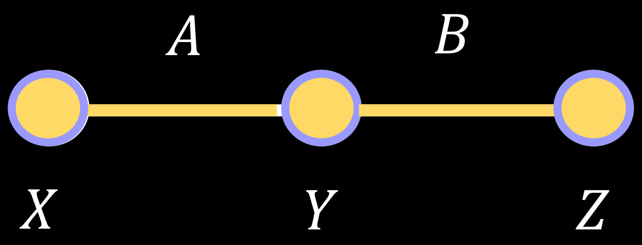


Figure 1.5

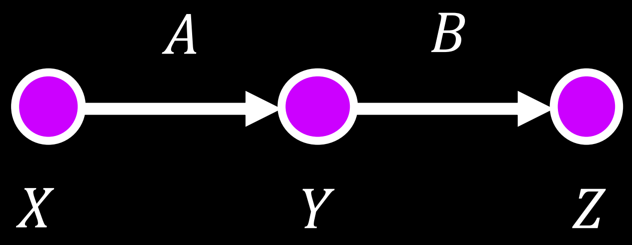
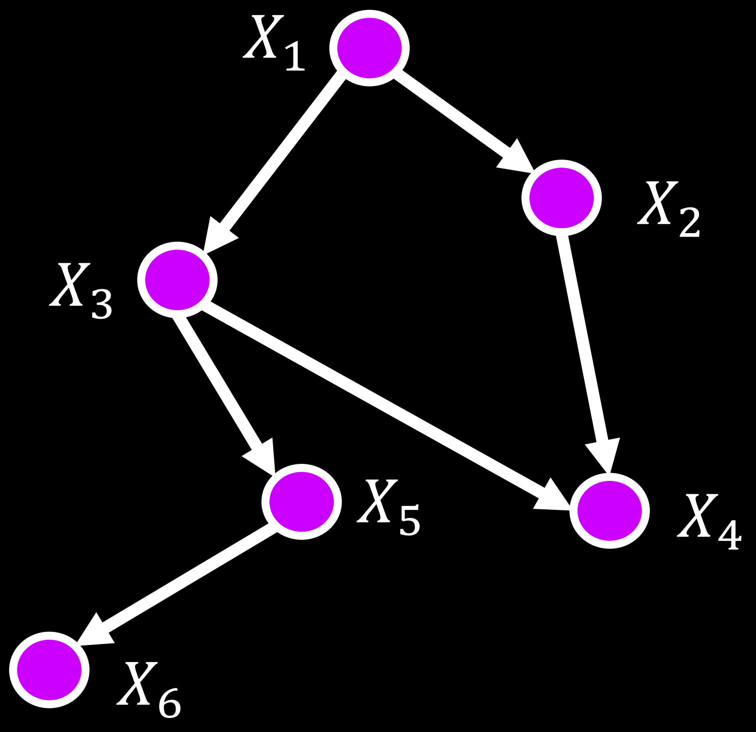
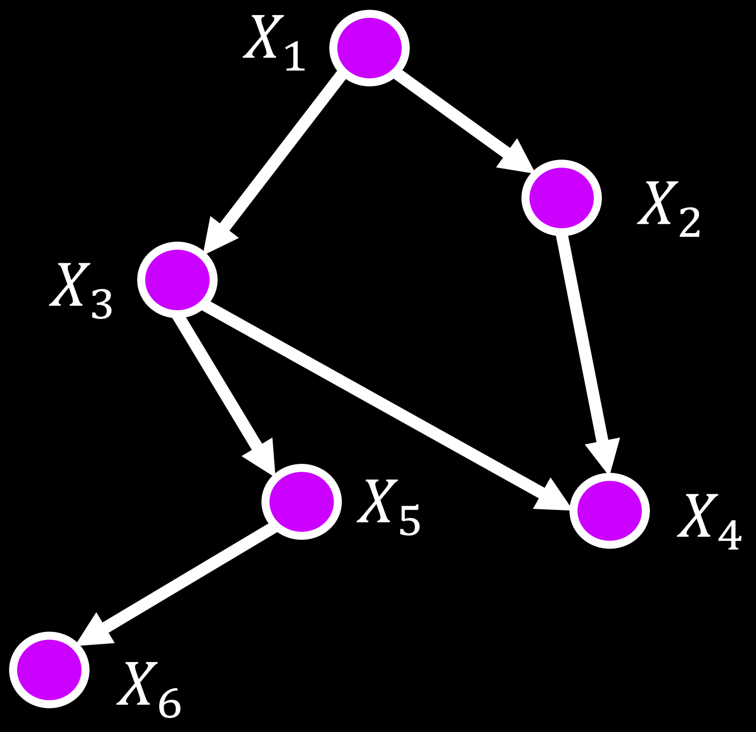


Figure 1.6

If **two nodes** are **connected by a directed path**, then **the first node is the ancestor** of every node in the path, and every node in the path is a **descendant of the first node**.



If **two nodes** are **connected by a directed path**, then **the first node is the ancestor** of every node in the path, and every node in the path is a **descendant of the first node**.



- $an(X_1) = \{\emptyset\}$; $an(X_2) = \{X_1\}$; $an(X_3) = \{X_1\}$;
- $an(X_4) = \{X_1, X_2, X_3\}$; $an(X_5) = \{X_1, X_3\}$; $an(X_6) = \{X_1, X_3, X_5\}$;

- $de(X_1) = \{X_2, X_3, X_4, X_5, X_6\}$; $de(X_2) = \{X_4\}$; $de(X_3) = \{X_4, X_5, X_6\}$;
- $de(X_4) = \{\emptyset\}$; $de(X_5) = \{X_6\}$; $de(X_6) = \{\emptyset\}$

When a **directed path** exists from a node to itself, the path is called **cyclic**.

A **directed graph** without cycles is an **acyclic graph**.

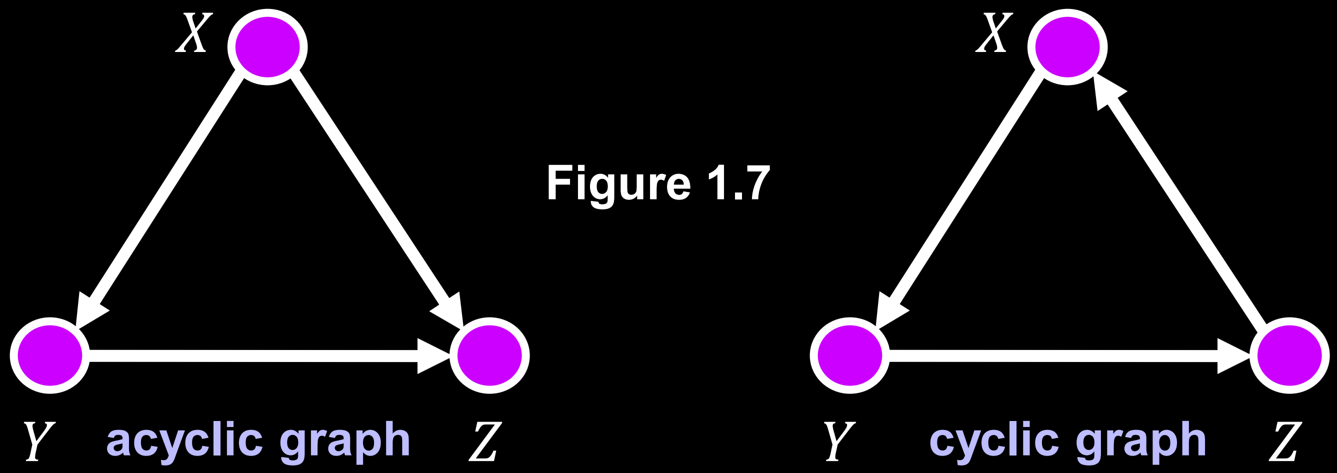


Figure 1.7

In order to deal rigorously with questions of causality, we must have a way to formally setting down our assumptions about the causal story behind a data set.

We introduce the **Structural Causal Model** (**SCM**), which is used to describe the relevant features of the world and how they interact with each other.

A **Structural Causal Model** describes how nature assigns values to variables of interest.

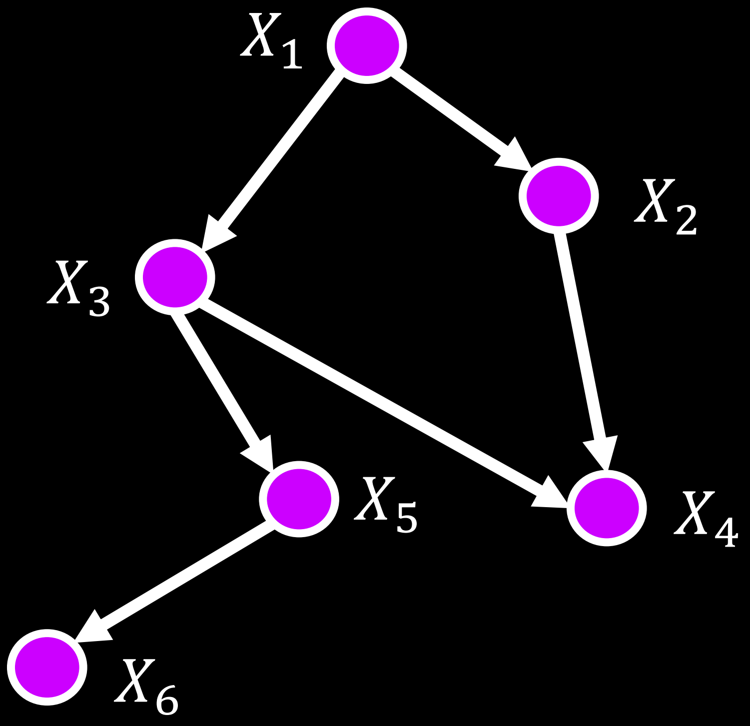
set of **exogenous variables**	set of **endogenous variables**	set of **functions** on endogenous variables
U	V	F

Each function $f \in F$ assigns each variable in V a value based on the values of the other variables in the model.

A variable *X* is a **direct cause** of a variable *Y*, if *X* appears in the function that assigns *Y*'s value.

A variable *X* is a **cause** of a variable *Y* if *X* is a **direct cause** of *Y*, or a **cause of any cause** of *Y*.

A variable *Y* is a **potential cause** of *X*, if *X* is a descendant of *Y*.



X_1 is a **direct cause** of X_2, X_3

X_2 is a **direct cause** of X_4

X_3 is a **direct cause** of X_4, X_5

X_5 is a **direct cause** of X_6

X_1 is a **cause** (potential cause) of X_2, X_3, X_4, X_5 , X_6

X_2 is a **cause** (potential cause) of X_4

X_3 is a **cause** (potential cause) of X_4, X_5, X_6

X_5 is a **cause** (potential cause) of X_6

set of

exogenous variables

U

Exogenous variables can not be descendant of any other variables, and in particular, can not be descendant of an endogenous variable; they have no ancestors and are represented as root nodes in graphs.

They are external to the model; we chose, for whatever reason, not to explain how they are caused set of

endogenous

variables

V

Every endogenous variable in a model is descendant of at least one exogeneous variable.

If we know the value of every exogenous variable, then using functions in F, we can determine with perfect certainty the value of every endogenous variable.

Suppose we are interested in studying the causal relationships between a

- **treatment *X*** and,
- **lung function *Y***

for individuals who suffer from **asthma**.

We might assume that *Y* also depends on, or is caused by, **air pollution levels** as captured by a variable ***Z***.

- *X* and *Y* are endogenous
- *Z* is exogenous

this is because we assume that air pollution is an external factor, that is, it can not be caused by an individual's selected treatment or their lung function.

Every SCM is associated with a **Graphical Causal Model**, referred to informally as a “**Graphical Model**” or simply a “**Graph**”.

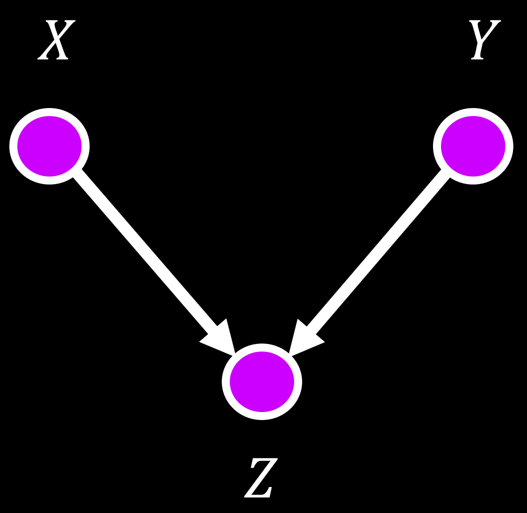
SCM 1.5.1 (**Salary Based on Education and Experience**)

X years of schooling

Y years of employment

Z salary

Figure 1.9



$U = \{X, Y\} \quad V = \{Z\} \quad F = \{f_Z\}$

$f_Z: Z = 2\,X + 3\,Y$

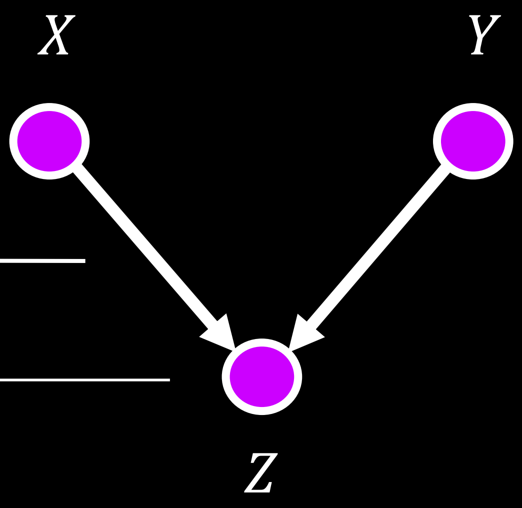
$M = \langle U, V, F \rangle \longrightarrow G = \langle (U, V), E \rangle$

1.5 Structural Causal Models: Modeling Causal Assumptions

Every SCM is associated with a **Graphical Causal Model**, referred to informally as a "**Graphical Model**" or simply a "**Graph**".

SCM 1.5.1 (Salary Based on Education and Experience)

Figure 1.9



$U = \{X, Y\}$

$f_Z : Z = f_Z(X, Y)$

$F = \{f_Z\} \leftarrow V = \{Z\}$

$???$

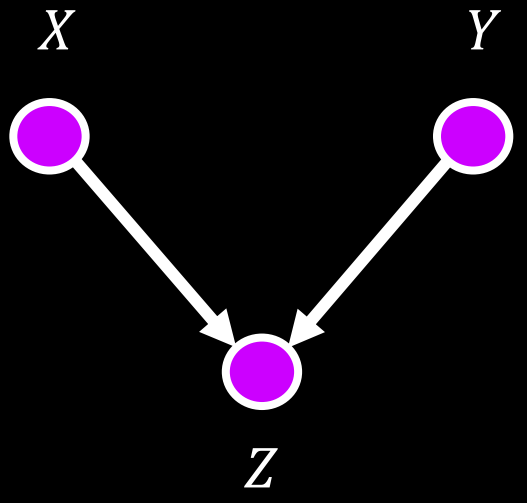
How *X* and *Y* cause *Z*? $\longrightarrow M = \langle U, V, ???\rangle$ $\quad M = \langle U, V, F\rangle \longleftarrow G = \langle (U, V), E\rangle$

Every SCM is associated with a **Graphical Causal Model**, referred to informally as a "**Graphical Model**" or simply a "**Graph**".

If graphical models contain less information than SCMs, why do we use them at all?

- The **knowledge** that we have about **causal relationships** is **not quantitative**, as demanded by SCM, but **qualitative**, as represented in a graphical model.

Figure 1.9



$$M = \langle U, V, ???\rangle \longleftarrow G = \langle (U, V), E\rangle$$

We know off-hand that **sex** is a cause of **height** and that **height and sex** are causes of **performance** in basketball, but we would hesitate to give numerical values to these relationships.

We could, instead of drawing a graph, simply create a **partially specified version of the SCM**.

$$M = \langle U, V, ???\rangle$$

SCM 1.5.2 (**Basketball Performance based on Height and Sex**)

$$V = \{Height, Sex, Performance\}$$

$$U = \{U_1, U_2, U_3\}$$

Error Terms (or omitted factors) →

Unmeasured factors that we do not care to mention but that affect the variables in V that we can measure.

Additional unknown and/or **random exogenous causes** of what we observe.

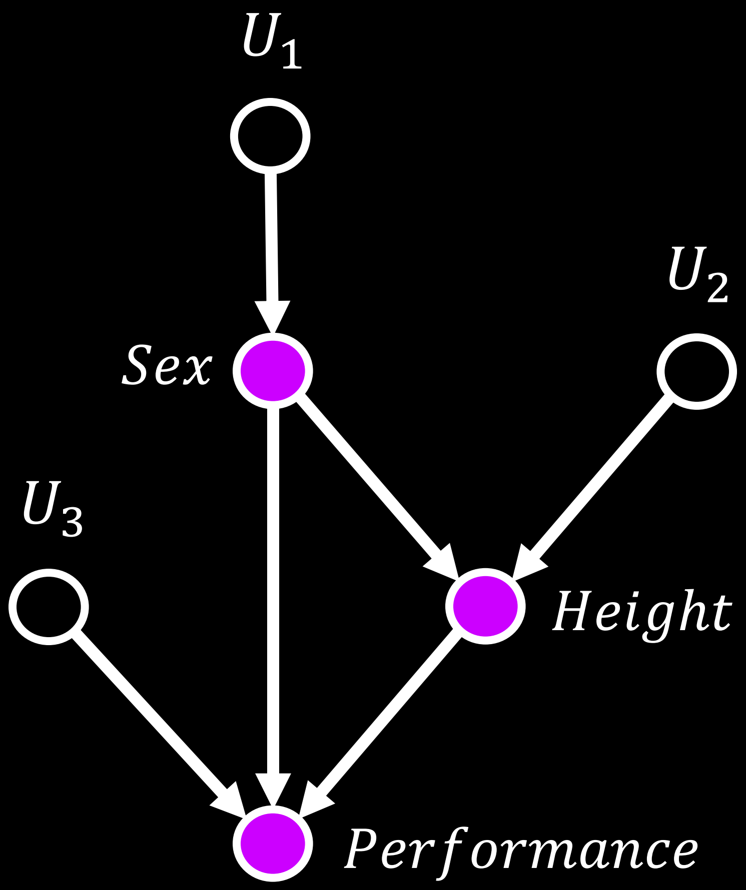
$$F = \{f_1, f_2\}$$

$$Sex = U_1 \qquad Height = f_1(Sex, U_2) \qquad Performance = f_2(Height, Sex, U_3)$$

We know off-hand that **sex** is a cause of **height** and that **height and sex** are causes of **performance** in basketball, but we would hesitate to give numerical values to these relationships.

We could, instead of drawing a graph, simply create a **partially specified version of the SCM**.

$$M = \langle U, V, ???\rangle$$



SCM 1.5.2 (Basketball Performance based on Height and Sex)

$$V = \{Height, Sex, Performance\}$$

$$U = \{U_1, U_2, U_3\}$$

$$F = \{f_1, f_2\}$$

$$Sex = U_1 \qquad Height = f_1(Sex, U_2) \qquad Performance = f_2(Height, Sex, U_3)$$

We know off-hand that **sex** is a cause of **height** and that **height and sex** are causes of **performance** in basketball, but we would hesitate to give numerical values to these relationships.

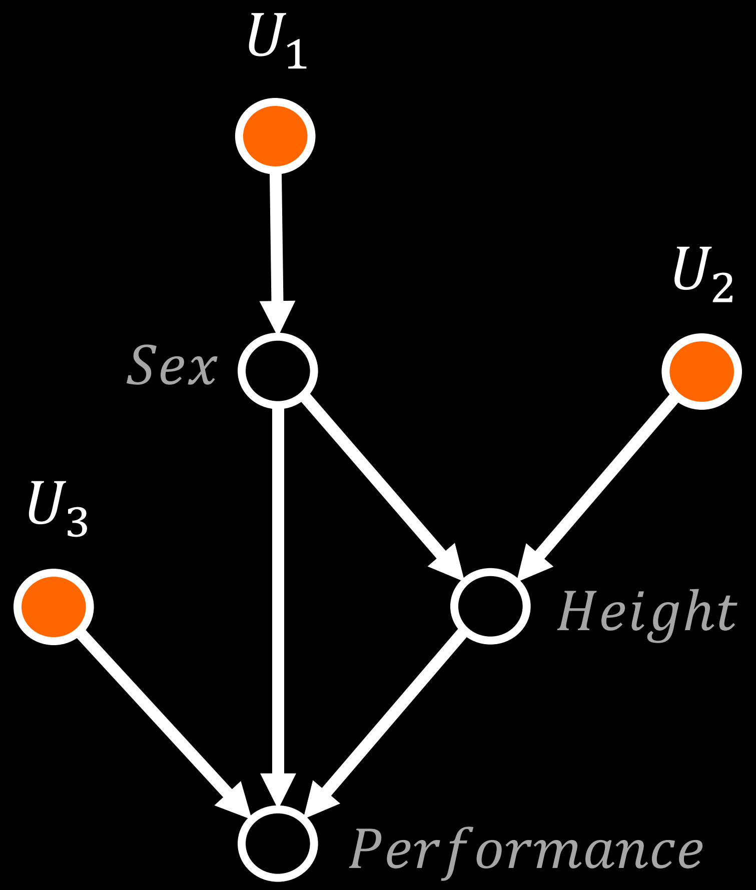
We could, instead of drawing a graph, simply create a **partially specified version of the SCM**.

$$M = \langle U, V, ???\rangle$$

SCM 1.5.2 (Basketball Performance based on Height and Sex)

$$U = \{U_1, U_2, U_3\}$$

Exogenous Variables



We know off-hand that **sex** is a cause of **height** and that **height and sex** are causes of **performance** in basketball, but we would hesitate to give numerical values to these relationships.

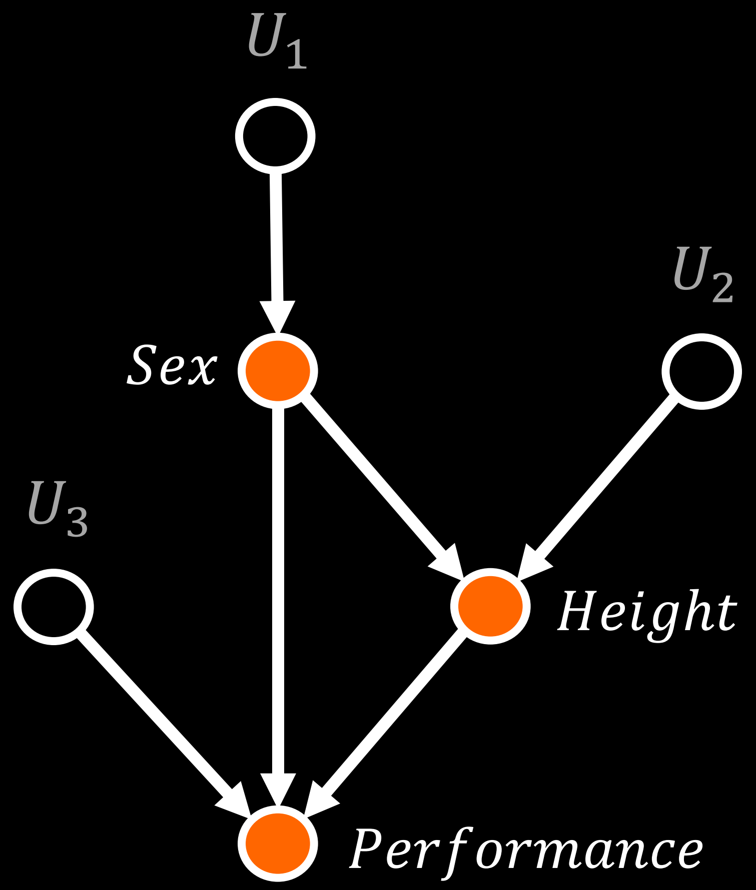
We could, instead of drawing a graph, simply create a **partially specified version of the SCM**.

$$M = \langle U, V, ???\rangle$$

SCM 1.5.2 (Basketball Performance based on Height and Sex)

$$V = \{Height, Sex, Performance\}$$

Endogenous Variables



Another advantage of Graphical Models is that they allow to express joint distributions very efficiently.

So far, we have presented **joint distributions** in two ways

Joint Probability Table

	Drug	No Drug
recovered	0.4	0.1
not recovered	0.2	0.3

$$2^2 = 4$$

Treatment = {Drug, No Drug}

Patient's Status = {recovered, not recovered}

10 binary variables require to specify

$$2^{10} = 1{,}024$$

probability values.

Fully specified SCM

Great efficiency: we need to specify the "*n*" functions that govern the relationships between the variables, and then from the probabilities of the **error terms**, we can discover all the probabilities that govern the joint probability distribution.

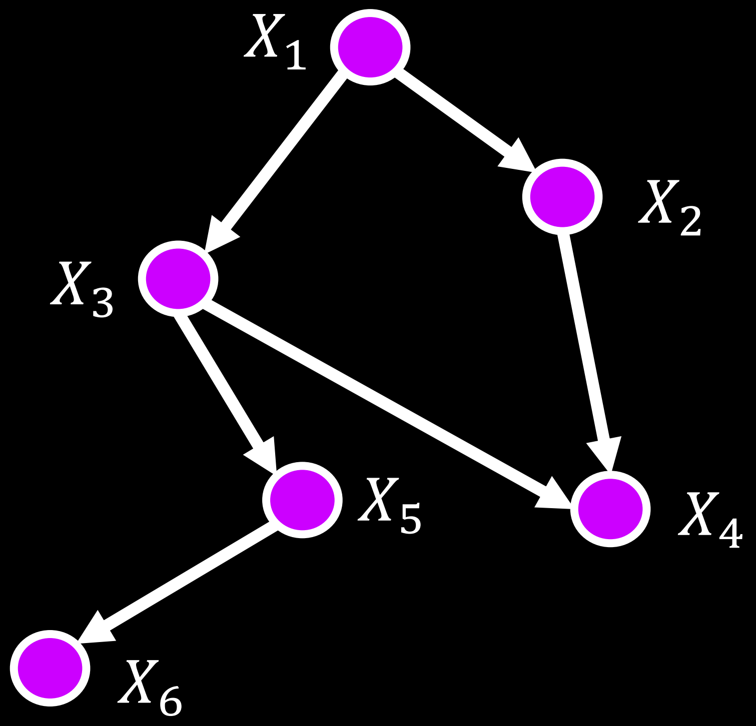
We are not always in a position to fully specify a SCM model *M*:

- We know a variable is a cause of another but we do not know the equation relating them
- We do not know the distribution of the error terms

Even if we know these objects, writing them down may be easier said than done, especially, when the variables are discrete and the functions do not have familiar algebraic expressions.

Graphical models help to overcome both previous barriers through the **Rule of Product Decomposition**.

For any model whose graph is acyclic, the joint distribution of the variables of the model is given by the product of the **conditional distributions** over all **families** in the graph.



All binary variables for simplicity.

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P\big(X_i | pa(X_i)\big)$$

$$P(X_1, X_2, X_3, X_4, X_5, X_6) = P(X_1)\, P(X_2|X_1)\, P(X_3|X_1)\, P(X_4|X_2, X_3)\, P(X_5|X_3)\, P(X_6|X_5)$$

$$2 + 2^2 + 2^2 + 2^3 + 2^2 + 2^2 = 26$$

$$2^6 = 64$$

Graphical models help to overcome both previous barriers through the **Rule of Product Decomposition**.

For any model whose graph is acyclic, the joint distribution of the variables of the model is given by the product of the **conditional distributions** over all **families** in the graph.

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid pa(X_i)\right)$$

Advantages of the graph representation

- saves a great deal of processing time in large models
- increases the accuracy of frequency counting

one high dimensional estimation problem

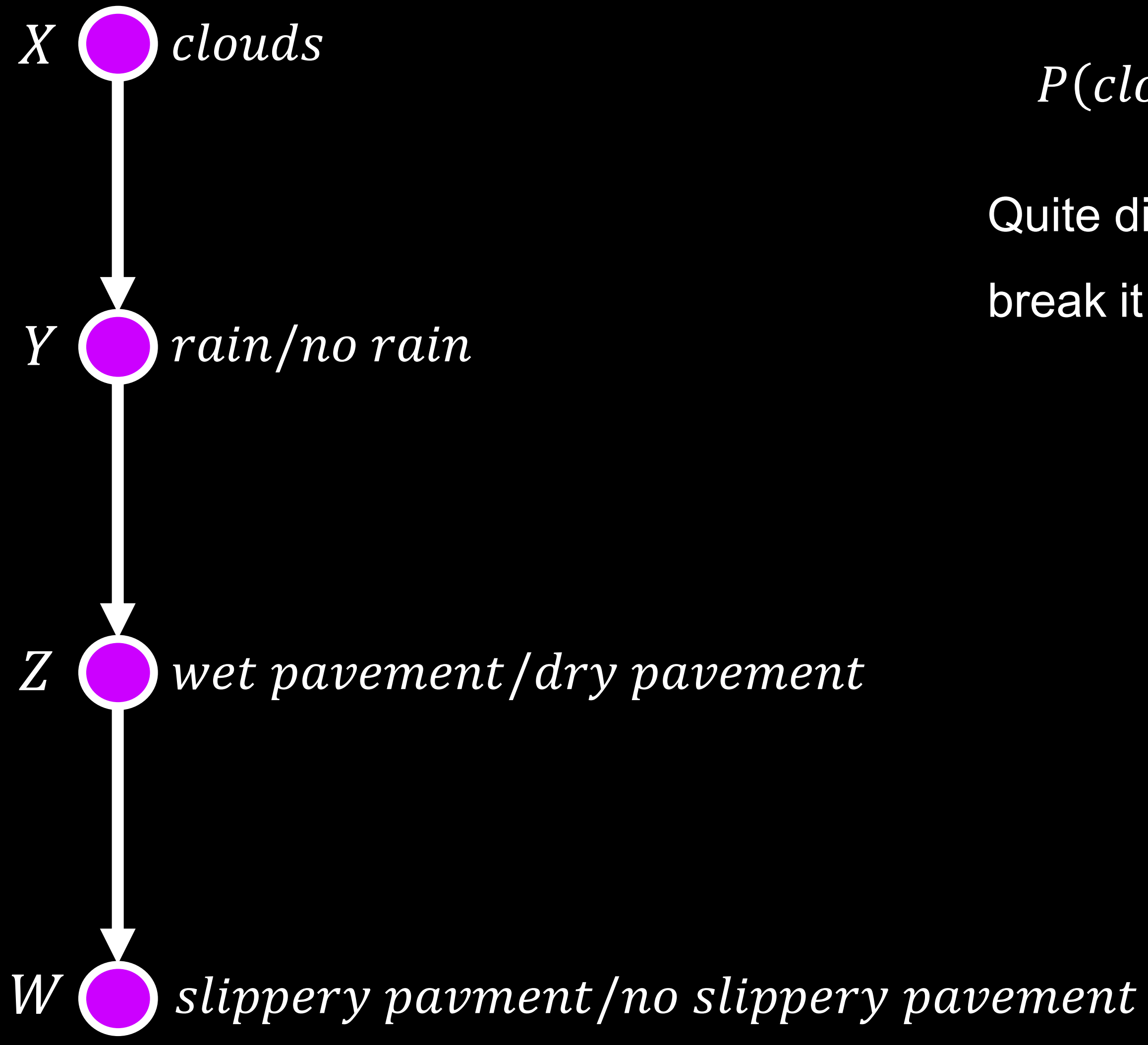
↓

the graph representation

↓

few low dimensional probability distribution challenges

1.5.2 Structural Causal Models: Product Decomposition



Based on your experience of the world, how plausible is that

$$P(clouds, no\ rain, dry\ pavement, slippery\ pavement) = 0.23$$

Quite difficult to answer, but using the **Product Rule**, we can break it into pieces

$P(clouds)$	0.5
$P(no\ rain \mid clouds)$	0.75
$P(dry\ pavement \mid no\ rain)$	0.9
$P(slippery\ pavement \mid dry\ pavement)$	0.05

$$0.5 \times 0.75 \times 0.9 \times 0.05 = 0.0169$$

The importance of the **Rule of Product Decomposition** is particularly appreciated when we deal with estimation.

A major problem is **effective sampling designs, and estimation strategies**, that allow us to exploit an appropriate data set to estimate the probabilities as precisely as we might need.

$P(clouds)$

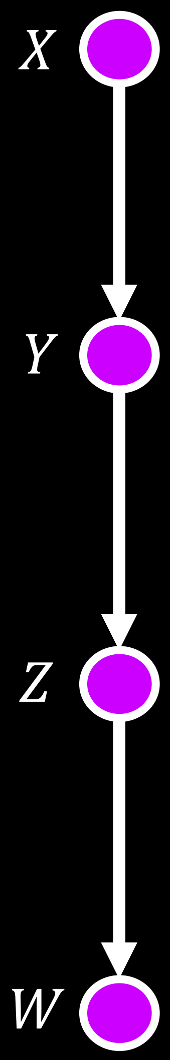
$P(no\ rain|clouds)$

$P(dry\ pavement|no\ rain)$

$P(slippery\ pavement|dry\ pavement)$

← estimate from data and not from our judgement.

number of combinations of X, Y, Z, W to be assigned probabilities is $2^4 - 1 = 15$



Assume your data set consists of 45 random observations, i.e. random assignments

(x, y, z, w)

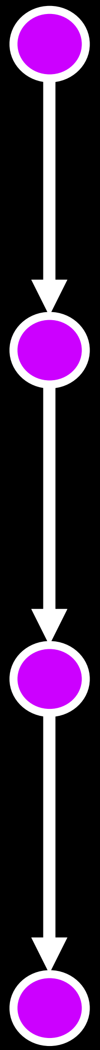
On the average, each random assignment would receive about 3 samples, i.e. 45/15.

However, some will receive 2, some 1 and some 0.

It is very unlikely that we would obtain a sufficient number of samples in each cell to assess the proportion in the population at large (i.e., when the sample size goes to infinity).

If we use our product rule, however, the 45 sample are separated into much larger categories.

number of combinations of X, Y, Z, W
to be assigned probabilities is $2^4 - 1 = 15$



X	*clouds*	$P(clouds)$	$\frac{45}{2} = 22.5$
Y	*rain/no rain*	$P(no\ rain \mid clouds)$	$\frac{45}{4} = 11.25$
Z	*wet pavement/dry pavement*	$P(dry\ pavement \mid no\ rain)$	$\frac{45}{4} = 11.25$
W	*slippery pavment/no slippery pavement*	$P(slippery\ pavement \mid dry\ pavement)$	$\frac{45}{4} = 11.25$