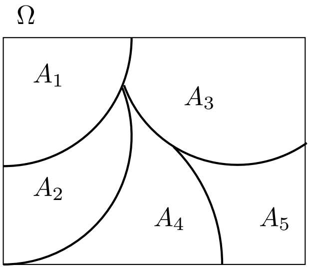
Axioms of Probability

Axioms of Probability

- For any event A,

$$P(A) \geq 0$$



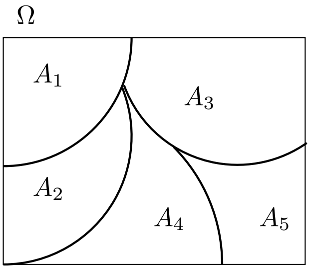
Axioms of Probability

- For any event A,

$$P(A) \geq 0$$

- If $A_1, A_2, A_3, ..., A_n$ are disjoint events (i.e., $A_i \cap A_j = \phi \quad \forall i \neq j$) then

$$P(\cup A_i) = \sum_i P(A_i)$$



Axioms of Probability

- For any event A,

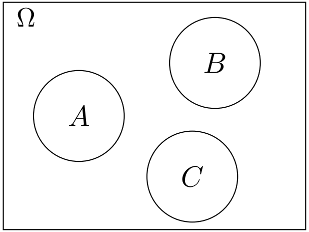
$$P(A) \geq 0$$

- If $A_1, A_2, A_3, ..., A_n$ are disjoint events (i.e., $A_i \cap A_j = \phi \quad \forall i \neq j$) then

$$P(\cup A_i) = \sum_i P(A_i)$$

- If Ω is the universal set containing all events then

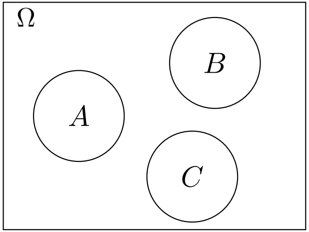
$$P(\Omega) = 1$$



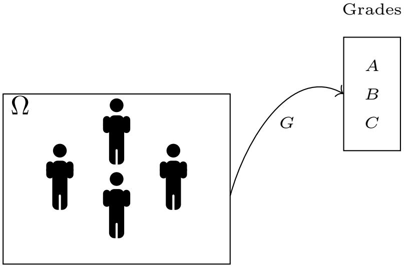
Random Variable (intuition)

- Suppose a student can get one of 3 possible grades in a course: A, B, C

Random Variable (intuition)

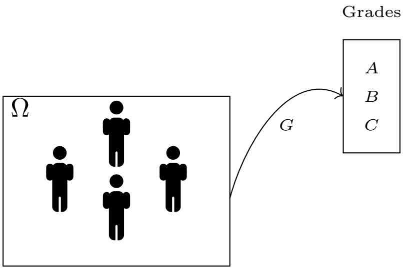


- Suppose a student can get one of 3 possible grades in a course: A, B, C
- One way of interpreting this is that there are 3 possible events here



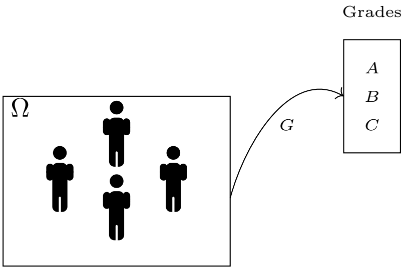
Random Variable (intuition)

- Suppose a student can get one of 3 possible grades in a course: A, B, C
- One way of interpreting this is that there are 3 possible events here
- Another way of looking at this is there is a *random variable* G which each student to one of the 3 possible values



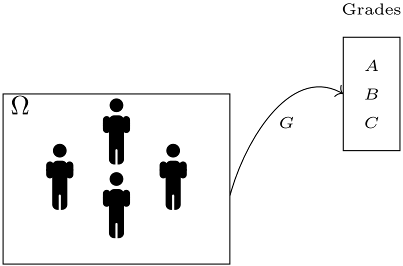
Random Variable (intuition)

- Suppose a student can get one of 3 possible grades in a course: A, B, C
- One way of interpreting this is that there are 3 possible events here
- Another way of looking at this is there is a *random variable* G which each student to one of the 3 possible values
- And we are interested in $P(G = g)$ where $g \in \{A, B, C\}$



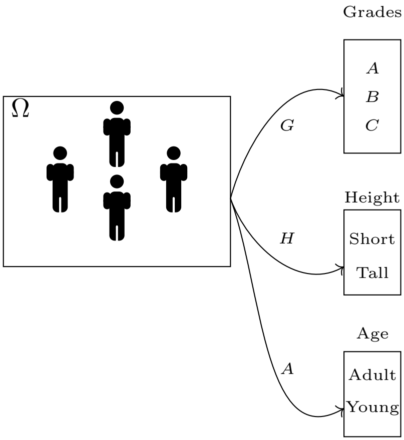
Random Variable (intuition)

- Suppose a student can get one of 3 possible grades in a course: A, B, C
- One way of interpreting this is that there are 3 possible events here
- Another way of looking at this is there is a *random variable* G which each student to one of the 3 possible values
- And we are interested in $P(G = g)$ where $g \in \{A, B, C\}$
- Of course, both interpretations are conceptually equivalent



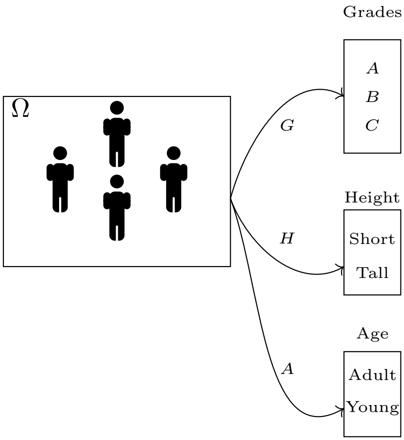
Random Variable (intuition)

- But the second one (using random variables) is more compact



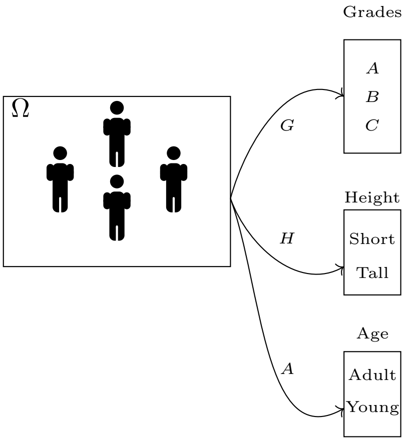
Random Variable (intuition)

- But the second one (using random variables) is more compact
- Specially, when there are multiple attributes associated with a student (outcome) - *grade, height, age, etc.*



Random Variable (intuition)

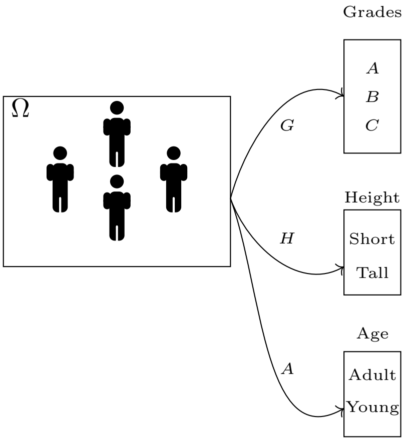
- But the second one (using random variables) is more compact
- Specially, when there are multiple attributes associated with a student (outcome) - *grade, height, age, etc.*
- We could have one random variable corresponding to each attribute

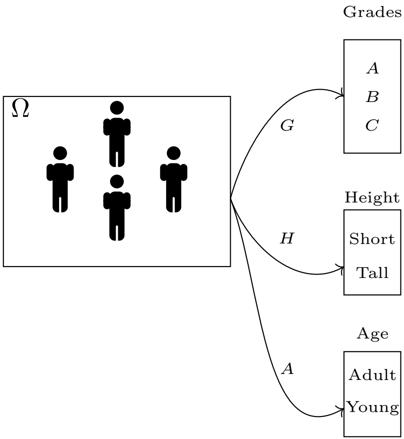


Random Variable (intuition)

- But the second one (using random variables) is more compact
- Specially, when there are multiple attributes associated with a student (outcome) - *grade, height, age, etc.*
- We could have one random variable corresponding to each attribute
- And then ask for outcomes (or students) where $Grade = g$, $Height = h$, $Age = a$ and so on

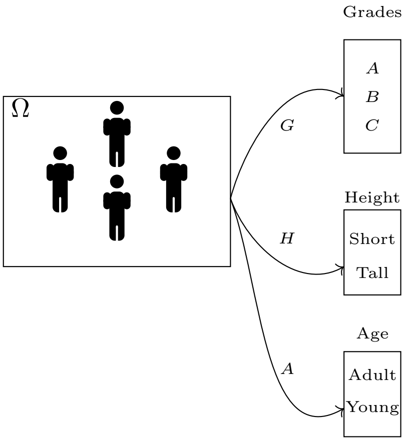
Random Variable (formal)





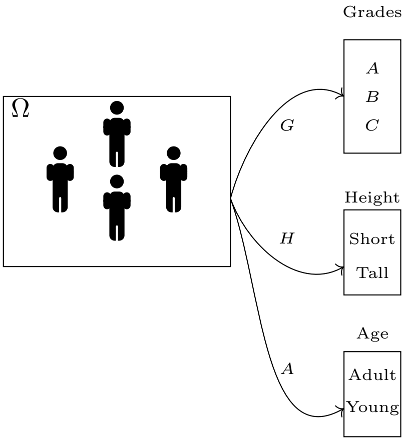
Random Variable (formal)

- A random variable is a ***function*** which maps each outcome in Ω to a value



Random Variable (formal)

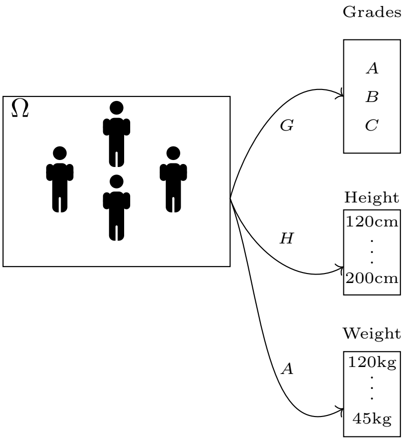
- A random variable is a ***function*** which maps each outcome in Ω to a value
- In the previous example, G (or f_{grade}) maps each student in Ω to a value: A, B or C



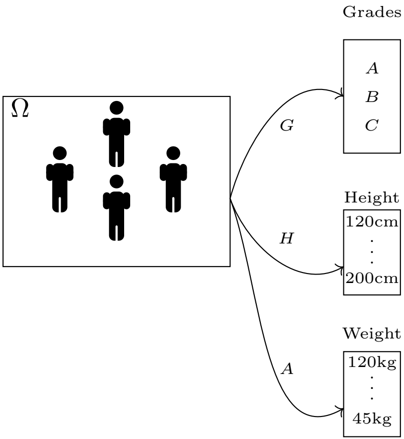
Random Variable (formal)

- A random variable is a ***function*** which maps each outcome in Ω to a value
- In the previous example, G (or f_{grade}) maps each student in Ω to a value: A, B or C
- The event $Grade = A$ is a shorthand for the event $\{\omega \in \Omega : f_{Grade} = A\}$

Random Variable (continuous v/s discrete)

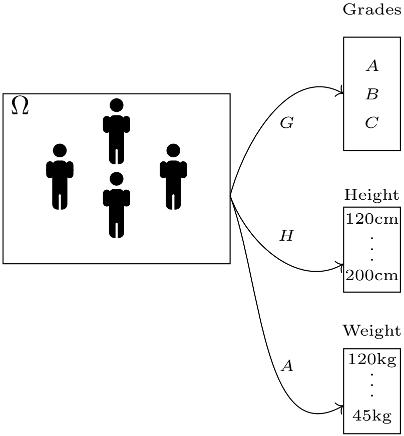


- A random variable can either take continuous values (for example, $weight, height$)



Random Variable (continuous v/s discrete)

- A random variable can either take continuous values (for example, $weight, height$)
- Or discrete values (for example, $grade, nationality$)



Random Variable (continuous v/s discrete)

- A random variable can either take continuous values (for example, *weight*, *height*)
- Or discrete values (for example, *grade*, *nationality*)
- For this discussion we will mainly focus on discrete random variables

Marginal Distribution

- What do we mean by ***marginal distribution*** over a random variable ?

Marginal Distribution

- What do we mean by ***marginal distribution*** over a random variable ?
- Consider our random variable G for grades

G	$P(G = g)$
A	0.1
B	0.2
C	0.7

Marginal Distribution

- What do we mean by ***marginal distribution*** over a random variable ?
- Consider our random variable G for grades
- Specifying the marginal distribution over G means specifying

$$P(G = g) \quad \forall g \in A, B, C$$

G	$P(G = g)$
A	0.1
B	0.2
C	0.7

Marginal Distribution

- What do we mean by ***marginal distribution*** over a random variable ?
- Consider our random variable G for grades
- Specifying the marginal distribution over G means specifying

$$P(G = g) \quad \forall g \in A, B, C$$

- We denote this marginal distribution compactly by $P(G)$

Joint Distribution

- Consider two random variable G (grade) and I (intellegence $\in \{$**H**igh, **L**ow$\}$)

Joint Distribution

- Consider two random variable G (grade) and I (intellegence $\in \{\textbf{H}\text{igh}, \textbf{L}\text{ow}\}$)
- The joint distribution over these two random variables assigns probabilities to all events involving these two random variables

$$P(G = g, I = i) \quad \forall (g, i) \in \{A, B, C\} \times \{H, L\}$$

Joint Distribution

G	I	$P(G = g, I = i)$
A	High	0.3
A	Low	0.1
B	High	0.15
B	Low	0.15
C	High	0.1
C	Low	0.2

- Consider two random variable G (grade) and I (intellegence $\in$ {**H**igh, **L**ow})
- The joint distribution over these two random variables assigns probabilities to all events involving these two random variables

$$P(G = g, I = i) \quad \forall (g, i) \in \{A, B, C\} \times \{H, L\}$$

Joint Distribution

G	I	$P(G = g, I = i)$
A	High	0.3
A	Low	0.1
B	High	0.15
B	Low	0.15
C	High	0.1
C	Low	0.2

- Consider two random variable G (grade) and I (intellegence $\in \{\textbf{H}\text{igh}, \textbf{L}\text{ow}\}$)
- The joint distribution over these two random variables assigns probabilities to all events involving these two random variables

$$P(G = g, I = i) \quad \forall (g, i) \in \{A, B, C\} \times \{H, L\}$$

- We denote this joint distribution compactly by $P(G, I)$

Conditional Distribution

G	$P(G\|I=H)$
A	0.6
B	0.3
C	0.1

G	$P(G\|I=L)$
A	0.3
B	0.4
C	0.3

- Consider two random variable G (grade) and I (intel-legence)

G	$P(G\|I = H)$
A	0.6
B	0.3
C	0.1

G	$P(G\|I = L)$
A	0.3
B	0.4
C	0.3

Conditional Distribution

- Consider two random variable G (grade) and I (intellegence)
- Suppose we are given the value of I (say, $I = H$) then the conditional distribution $P(G|I)$ is defined as

$$P(G = g|I = H) = \frac{P(G = g, I = H)}{P(I = H)} \forall g \in \{A, B, C\}$$

Conditional Distribution

G	$P(G\|I = H)$
A	0.6
B	0.3
C	0.1

G	$P(G\|I = L)$
A	0.3
B	0.4
C	0.3

- Consider two random variable G (grade) and I (intel-legence)
- Suppose we are given the value of I (say, $I = H$) then the conditional distribution $P(G|I)$ is defined as

$$P(G = g|I = H) = \frac{P(G = g, I = H)}{P(I = H)} \forall g \in \{A, B, C\}$$

- More compactly defined as

$$P(G|I) = \frac{P(G, I)}{P(I)}$$

$$or \quad \underbrace{P(G, I)}_{joint} = \underbrace{P(G|I)}_{conditional} * \underbrace{P(I)}_{marginal}$$

Joint Distribution (n random variables)

X_1	$\ldots$	X_n	$P(X_1, X_2, \ldots, X_n)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$

$$\sum = 1$$

- The joint distribution of n random variables assigns probabilities to all events involving the n random variables,

Joint Distribution (n random variables)

X_1	$\ldots$	X_n	$P(X_1, X_2, \ldots, X_n)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$

$$\sum = 1$$

- The joint distribution of n random variables assigns probabilities to all events involving the n random variables,
- In other words it assigns

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$$

for all possible values that variable X_i can take

Joint Distribution (n random variables)

X_1	$\ldots$	X_n	$P(X_1, X_2, \ldots, X_n)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$

$$\sum = 1$$

- The joint distribution of n random variables assigns probabilities to all events involving the n random variables,
- In other words it assigns

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$$

for all possible values that variable X_i can take

- If each random variable X_i can take two values then the joint distribution will assign probabilities to the 2^n possible events

Joint Distribution (n random variables)

X_1	$\dots$	X_n	$P(X_1, X_2, \dots, X_n)$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$

- The joint distribution over two random variables X_1 and X_2 can be written as,

$$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)$$

X_1	$\ldots$	X_n	$P(X_1, X_2, \ldots, X_n)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$

Joint Distribution (n random variables)

- The joint distribution over two random variables X_1 and X_2 can be written as,

  $$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)$$

- Similarly for n random variables

  $$P(X_1, X_2, ..., X_n)$$

X_1	$\dots$	X_n	$P(X_1, X_2, \dots, X_n)$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$

Joint Distribution (n random variables)

- The joint distribution over two random variables X_1 and X_2 can be written as,

  $$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)$$

- Similarly for n random variables

  $$P(X_1, X_2, ..., X_n)$$
  $$= P(X_2, ..., X_n|X_1)P(X_1)$$

X_1	$\dots$	X_n	$P(X_1, X_2, \dots, X_n)$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$

Joint Distribution (n random variables)

- The joint distribution over two random variables X_1 and X_2 can be written as,

$$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)$$

- Similarly for n random variables

$$P(X_1, X_2, ..., X_n)$$
$$= P(X_2, ..., X_n|X_1)P(X_1)$$
$$= P(X_3, ..., X_n|X_1, X_2)P(X_2|X_1)P(X_1)$$

X_1	$\ldots$	X_n	$P(X_1, X_2, \ldots, X_n)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$

Joint Distribution (n random variables)

- The joint distribution over two random variables X_1 and X_2 can be written as,

  $$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)$$

- Similarly for n random variables

  $$P(X_1, X_2, ..., X_n)$$
  $$= P(X_2, ..., X_n|X_1)P(X_1)$$
  $$= P(X_3, ..., X_n|X_1, X_2)P(X_2|X_1)P(X_1)$$
  $$= P(X_4, ..., X_n|X_1, X_2, X_3)P(X_3|X_2, X_1)$$
  $$P(X_2|X_1)P(X_1)$$

Joint Distribution (n random variables)

X_1	$\ldots$	X_n	$P(X_1, X_2, \ldots, X_n)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$

- The joint distribution over two random variables X_1 and X_2 can be written as,

$$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)$$

- Similarly for n random variables

$$\begin{aligned} & P(X_1, X_2, ..., X_n) \\ & = P(X_2, ..., X_n|X_1)P(X_1) \\ & = P(X_3, ..., X_n|X_1, X_2)P(X_2|X_1)P(X_1) \\ & = P(X_4, ..., X_n|X_1, X_2, X_3)P(X_3|X_2, X_1) \\ & \quad P(X_2|X_1)P(X_1) \\ & = P(X_1)\prod_{i=2}^{n} P(X_i|X_1^{i-1}) \quad \textbf{\textit{(chain rule)}} \end{aligned}$$

From Joint Distributions to Marginal Distributions

A	B	$P(A=a,B=b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A=a)$
High	0.55
Low	0.45

B	$P(B=a)$
High	0.65
Low	0.35

- Suppose we are given a joint distribtion over two random variables A, B

From Joint Distributions to Marginal Distributions

A	B	$P(A = a, B = b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A = a)$
High	0.55
Low	0.45

B	$P(B = a)$
High	0.65
Low	0.35

- Suppose we are given a joint distribtion over two random variables A, B
- The marginal distributions of A and B can be computed as

$$P(A = a) = \sum_{\forall b} P(A = a, B = b)$$

$$P(B = b) = \sum_{\forall a} P(A = a, B = b)$$

From Joint Distributions to Marginal Distributions

A	B	$P(A=a, B=b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A=a)$
High	0.55
Low	0.45

B	$P(B=a)$
High	0.65
Low	0.35

- Suppose we are given a joint distribtion over two random variables A, B
- The marginal distributions of A and B can be computed as

$$P(A=a) = \sum_{\forall b} P(A=a, B=b)$$

$$P(B=b) = \sum_{\forall a} P(A=a, B=b)$$

- More compactly written as

$$P(A) = \sum_{B} P(A, B)$$

$$P(B) = \sum_{A} P(A, B)$$

A	B	$P(A=a,B=b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A=a)$
High	0.55
Low	0.45

B	$P(B=a)$
High	0.65
Low	0.35

What if there are n random variables ?

- Suppose we are given a joint distribtion over n random variables X_1, X_2, ..., X_n

A	B	$P(A = a, B = b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A = a)$
High	0.55
Low	0.45

B	$P(B = a)$
High	0.65
Low	0.35

What if there are n random variables ?

- Suppose we are given a joint distribtion over n random variables X_1, X_2, ..., X_n
- The marginal distributions over X_1 can be computed as

$$P(X_1 = x_1)$$

$$= \sum_{\forall x_2, x_3, \ldots, x_n} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

A	B	$P(A=a,B=b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A=a)$
High	0.55
Low	0.45

B	$P(B=a)$
High	0.65
Low	0.35

What if there are n random variables ?

- Suppose we are given a joint distribtion over n random variables X_1, X_2, ..., X_n
- The marginal distributions over X_1 can be computed as

$$P(X_1 = x_1)$$
$$= \sum_{\forall x_2, x_3, \ldots, x_n} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

- More compactly written as

$$P(X_1) = \sum_{X_2, X_3, \ldots, X_n} P(X_1, X_2, \ldots, X_n)$$

Conditional Independence

- Two random variables X and Y are said to be independent if

$$P(X|Y) = P(X)$$

Conditional Independence

- Two random variables X and Y are said to be independent if

$$P(X|Y) = P(X)$$

- We denote this as $X \perp\!\!\!\perp Y$

Conditional Independence

- Two random variables X and Y are said to be independent if

$$P(X|Y) = P(X)$$

- We denote this as $X \perp\!\!\!\perp Y$
- In other words, knowing the value of Y does not change our belief about X

Conditional Independence

- Two random variables X and Y are said to be independent if

$$P(X|Y) = P(X)$$

- We denote this as $X \perp\!\!\!\perp Y$
- In other words, knowing the value of Y does not change our belief about X
- We would expect $\boldsymbol{G}rade$ to be dependent on $\boldsymbol{I}ntelligence$ but independent of $\boldsymbol{W}eight$

- Recall that by Chain Rule of Probability

$$P(X,Y) = P(X)P(Y|X)$$

Conditional Independence

- Two random variables X and Y are said to be independent if

$$P(X|Y) = P(X)$$

- We denote this as $X \perp\!\!\!\perp Y$
- In other words, knowing the value of Y does not change our belief about X
- We would expect $\boldsymbol{G}rade$ to be dependent on $\boldsymbol{I}ntelligence$ but independent of $\boldsymbol{W}eight$

Conditional Independence

- Recall that by Chain Rule of Probability

$$P(X,Y) = P(X)P(Y|X)$$

- However, if X and Y are independent, then

$$P(X,Y) = P(X)P(Y)$$

- Two random variables X and Y are said to be independent if

$$P(X|Y) = P(X)$$

- We denote this as $X \perp\!\!\!\perp Y$
- In other words, knowing the value of Y does not change our belief about X
- We would expect $\boldsymbol{G}rade$ to be dependent on $\boldsymbol{I}ntelligence$ but independent of $\boldsymbol{W}eight$