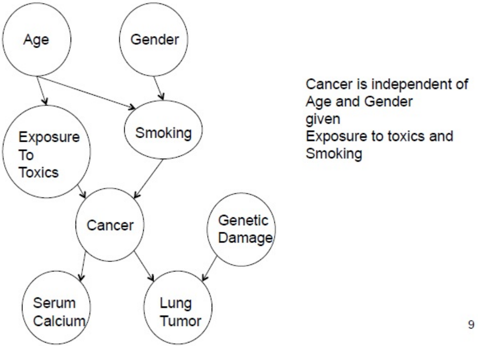
# Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 19
March 16, 2023

# Sampling methods



Cancer is independent of
Age and Gender
given
Exposure to toxics and
Smoking

# Sampling methods

Conditional independence Suppose $X, Y, Z$ are three rvs and the conditional distribution of $X$, given $Y$ and $Z$ does not depend on the value of $Y$ i.e.

$$p(X|\,Y,Z) = p(X|\,Z)$$

Then we say that $X$ is conditionally independent of $Y$ given $Z$.

# Sampling methods

Conditional independence Suppose $X, Y, Z$ are three rvs and the conditional distribution of $X$, given $Y$ and $Z$ does not depend on the value of $Y$ i.e.

$$p(X|\, Y, Z) = p(X|\, Z)$$

Then we say that $X$ is conditionally independent of $Y$ given $Z$.

Any example?

# Sampling methods

Conditional independence Suppose $X, Y, Z$ are three rvs and the conditional distribution of $X$, given $Y$ and $Z$ does not depend on the value of $Y$ i.e.

$$p(X|\,Y, Z) = p(X|\,Z)$$

Then we say that $X$ is conditionally independent of $Y$ given $Z$.
Any example?

In that case,

$$p(X, Y|Z) = p(X|Y, Z)\,p(Y|Z) = p(X|Z)\,p(Y|Z)$$

which means Conditioned on $Z$, the joint distribution of $X$ and $Y$ factorizes into product of the marginal distribution of $X$ and the marginal distribution of $Y$ (both conditioned on $Z$)

# Sampling methods

Conditional independence Suppose $X, Y, Z$ are three rvs and the conditional distribution of $X$, given $Y$ and $Z$ does not depend on the value of $Y$ i.e.

$$p(X|\,Y, Z) = p(X|\,Z)$$

Then we say that $X$ is conditionally independent of $Y$ given $Z$.
Any example?

In that case,

$$p(X, Y|Z) = p(X|Y, Z)\,p(Y|Z) = p(X|Z)\,p(Y|Z)$$

which means Conditioned on $Z$, the joint distribution of $X$ and $Y$ factorizes into product of the marginal distribution of $X$ and the marginal distribution of $Y$ (both conditioned on $Z$) This further means $X$ and $Y$ are statistically independent given $Z$. We denote it as

$$X \perp\!\!\!\perp Y \,|\, Z$$

# Sampling methods

Question Do you see any advantage using it for joint distribution associated with graphical models?

# Sampling methods

Question Do you see any advantage using it for joint distribution associated with graphical models? Does structural properties of the graph reflect conditional independence ?

# Sampling methods

Question Do you see any advantage using it for joint distribution associated with graphical models? Does structural properties of the graph reflect conditional independence ?

Consider the model:

# Sampling methods

Question Do you see any advantage using it for joint distribution associated with graphical models? Does structural properties of the graph reflect conditional independence ?

Consider the model:
Then $P(X, Y, Z) = p(X|Z)p(Y|Z)p(Z)$

# Sampling methods

Question Do you see any advantage using it for joint distribution associated with graphical models? Does structural properties of the graph reflect conditional independence ?

Consider the model:
Then $P(X, Y, Z) = p(X|Z)p(Y|Z)p(Z)$ if none of the variables are observed then we can investigate whether $X$ and $Y$ are independent with respect to $Z$ :

# Sampling methods

Question Do you see any advantage using it for joint distribution associated with graphical models? Does structural properties of the graph reflect conditional independence ?

Consider the model:
Then $P(X, Y, Z) = p(X|Z)p(Y|Z)p(Z)$ if none of the variables are observed then we can investigate whether $X$ and $Y$ are independent with respect to $Z$ :

$$p(X, Y) = \sum_{z} p(x, y, z) = \sum_{z} p(x|z)p(y|z)p(z)$$

so it does not factorize into the product $p(x)p(y)$, and so

$$X \not\perp\!\!\!\perp Y \,|\emptyset,$$

in general, where $\emptyset$ denotes the empty set

# Sampling methods

Now if the value of $z$ is observed then

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x|z)p(y|z)p(z)}{p(z)} = p(x|z)p(y|z)$$

# Sampling methods

Now if the value of $z$ is observed then

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x|z)p(y|z)p(z)}{p(z)} = p(x|z)p(y|z)$$

Question What is the observation from the graph?

# Sampling methods

Now if the value of $z$ is observed then

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x|z)p(y|z)p(z)}{p(z)} = p(x|z)p(y|z)$$

Question What is the observation from the graph?

→ Note that the graph is a path

# Sampling methods

Now if the value of $z$ is observed then

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x|z)p(y|z)p(z)}{p(z)} = p(x|z)p(y|z)$$

Question What is the observation from the graph?

- → Note that the graph is a path
- → The node corresponding to $Z$ is said to be tail-to-tail with respect to this path because the node is connected to the tails of the two directions, and the presence of such a path connecting $X$ and $Y$ causes these to be independent

# Sampling methods

Now if the value of $z$ is observed then

$$p(x,y|z) = \frac{p(x,y,z)}{p(z)} = \frac{p(x|z)p(y|z)p(z)}{p(z)} = p(x|z)p(y|z)$$

Question What is the observation from the graph?

- → Note that the graph is a path
- → The node corresponding to $Z$ is said to be tail-to-tail with respect to this path because the node is connected to the tails of the two directions, and the presence of such a path connecting $X$ and $Y$ causes these to be independent
- → Thus the conditioned node 'blocks' the path from $X$ to $Y$ and causes $X$ and $Y$ to be conditionally independent

# Sampling methods

Consider another graphical model whose joint distribution is

$$p(x, y, x) = p(x)p(z|x)p(y|z)$$

and suppose that none of the variables are observed.

# Sampling methods

Consider another graphical model whose joint distribution is

$$p(x, y, x) = p(x)p(z|x)p(y|z)$$

and suppose that none of the variables are observed. Then

$$p(x, y) = \sum_{z} p(x, y, z) = p(x) \sum_{z} p(z|x)p(y|z)$$

which does not factorize into product $p(x)p(y)$, and hence

$$X \not\perp\!\!\!\perp Y \,|\emptyset$$

# Sampling methods

Consider another graphical model whose joint distribution is

$$p(x, y, x) = p(x)p(z|x)p(y|z)$$

and suppose that none of the variables are observed. Then

$$p(x, y) = \sum_z p(x, y, z) = p(x) \sum_z p(z|x)p(y|z)$$

which does not factorize into product $p(x)p(y)$, and hence

$$X \not\perp\!\!\!\perp Y \,|\emptyset$$

Next suppose $Z$ is observed, and we condition on $Z$ then

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x)p(z|x)p(y|z)}{p(z)} = p(x|z)p(y|z)$$

i.e.

$$X \perp\!\!\!\perp Y \,|Z$$

# Sampling methods

Question What is the observation from the graph?

# Sampling methods

Question What is the observation from the graph?

The node $Z$ is said to be head-to-tail wrt the path from $X$ to $Y$

# Sampling methods

Question What is the observation from the graph?

The node $Z$ is said to be head-to-tail wrt the path from $X$ to $Y$

The path from $X$ to $Y$ is 'blocked' by $Z$ and we obtain conditional independence

# Sampling methods

Question What is the observation from the graph?

The node $Z$ is said to be head-to-tail wrt the path from $X$ to $Y$

The path from $X$ to $Y$ is 'blocked' by $Z$ and we obtain conditional independence

Another example Suppose the joint distribution of a graphical model is

$$p(x, y, z) = p(x)p(y)p(z|x, y)$$

# Sampling methods

Question What is the observation from the graph?

The node $Z$ is said to be head-to-tail wrt the path from $X$ to $Y$

The path from $X$ to $Y$ is 'blocked' by $Z$ and we obtain conditional independence

Another example Suppose the joint distribution of a graphical model is

$$p(x, y, z) = p(x)p(y)p(z|x, y)$$

Suppose none of the variables are observed. Then

$$p(x, y) = \sum_z p(x, y, z) = p(x)p(y) \sum_z p(z|x, y) = p(x)p(y),$$

so $X, Y$ are independent i.e. $X \perp\!\!\!\perp Y \,|\emptyset$ even if no variables are observed

# Sampling methods

Suppose we condition it on $Z$. Then

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x)p(y)p(z|x, y)}{p(z)}$$

which in general does not factorize into the product $p(x)p(y)$ and hence

$$X \not\perp\!\!\!\perp Y \,|\emptyset$$

# Sampling methods

Suppose we condition it on $Z$. Then

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x)p(y)p(z|x, y)}{p(z)}$$

which in general does not factorize into the product $p(x)p(y)$ and hence

$$X \not\perp\!\!\!\perp Y \,|\emptyset$$

Question What is the observation from the graph?

# Sampling methods

Suppose we condition it on $Z$. Then

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x)p(y)p(z|x, y)}{p(z)}$$

which in general does not factorize into the product $p(x)p(y)$ and hence

$$X \not\perp\!\!\!\perp Y \,|\emptyset$$

Question What is the observation from the graph?

The node $Z$ is head-to-head wrt to the path from $X$ to $Y$

# Sampling methods

Suppose we condition it on $Z$. Then

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x)p(y)p(z|x, y)}{p(z)}$$

which in general does not factorize into the product $p(x)p(y)$ and hence

$$X \not\perp\!\!\!\perp Y \,|\emptyset$$

Question What is the observation from the graph?

The node $Z$ is head-to-head wrt to the path from $X$ to $Y$

When the node $Z$ is unobserved, it ‘blocks’ the path and the variables $X, Y$ are independent

# Sampling methods

Suppose we condition it on $Z$. Then

$$p(x,y|z) = \frac{p(x,y,z)}{p(z)} = \frac{p(x)p(y)p(z|x,y)}{p(z)}$$

which in general does not factorize into the product $p(x)p(y)$ and hence

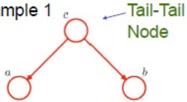
$$X \not\perp\!\!\!\perp Y\,|\emptyset$$

Question What is the observation from the graph?

- The node $Z$ is head-to-head wrt to the path from $X$ to $Y$
- When the node $Z$ is unobserved, it 'blocks' the path and the variables $X, Y$ are independent
- When $Z$ is observes, it 'unblocks' the path and renders $X, Y$ dependent

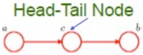
# Sampling methods

Example 1



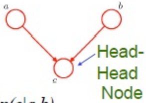
$p(a,b,c)=p(a|c)p(b|c)p(c)$

Example 2



$p(a,b,c)=p(a)p(c|a)p(b|c)$

Example 3



$p(a,b,c)=p(a)p(b)p(c|a,b)$