Basic math/stats review

Nisheeth

Overview

- Probability
  - Random variables, expected value
  - Common distributions, sufficient statistics
  - Conditional, marginal and joint distributions
  - Bayes rule
- Correlations
  - Linear correlations
  - Rank correlations
  - Entropy, mutual information
- Hypothesis testing
  - Basic tests
  - Cautions
  - Bayes Factors
- Inference
  - Estimation
  - Conjugacy
  - Applications

Introduction to Probability

Bonus question

Random Variable

- A random variable *x* takes on a defined set of values with different probabilities.
- Roughly, <u>probability</u> is how frequently we expect different outcomes to occur if we repeat the experiment over and over ("frequentist" view)

Random variables can be discrete or continuous

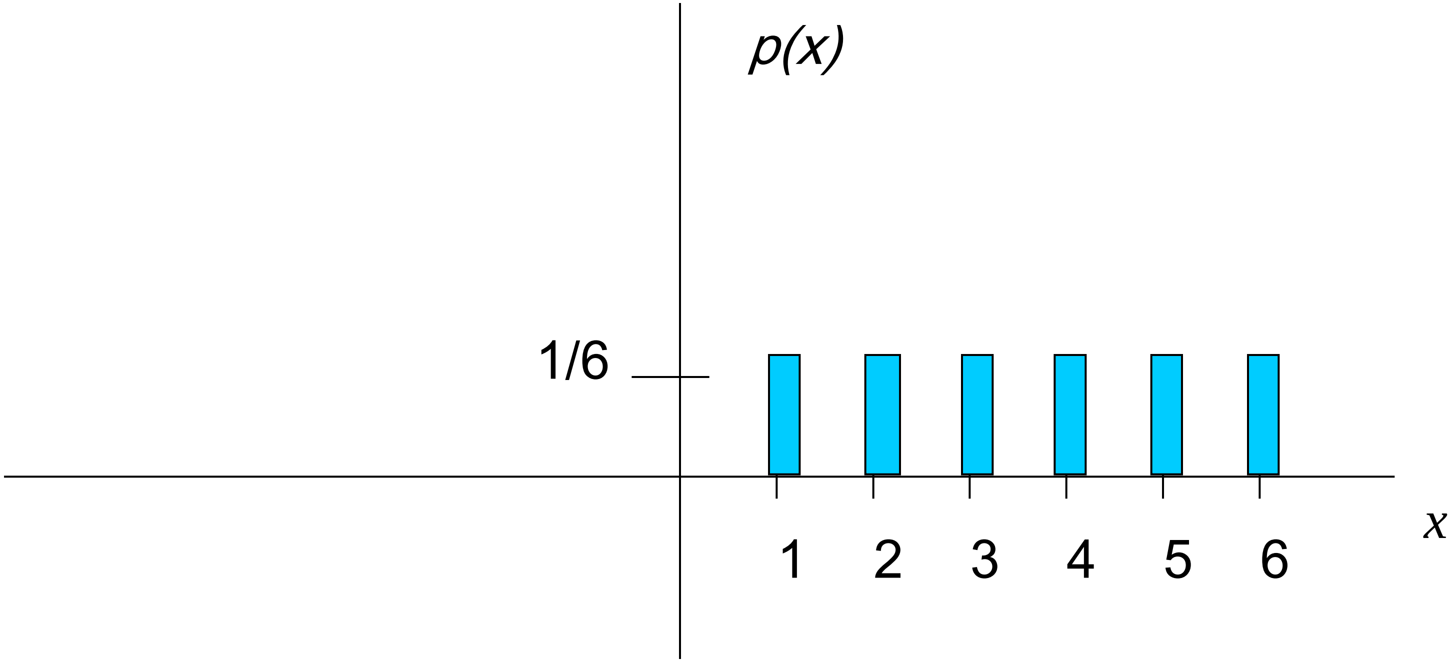
- **Discrete** random variables have a countable number of outcomes

- **Continuous** random variables have an infinite continuum of possible values.

Probability functions

- A probability function maps the possible values of *x* against their respective probabilities of occurrence, *p(x)*
- *p(x)* is a number from 0 to 1.0.
- The area under a probability function is always 1.

Discrete example: roll of a die



$$\sum_{\text{all } x} P(x) = 1$$

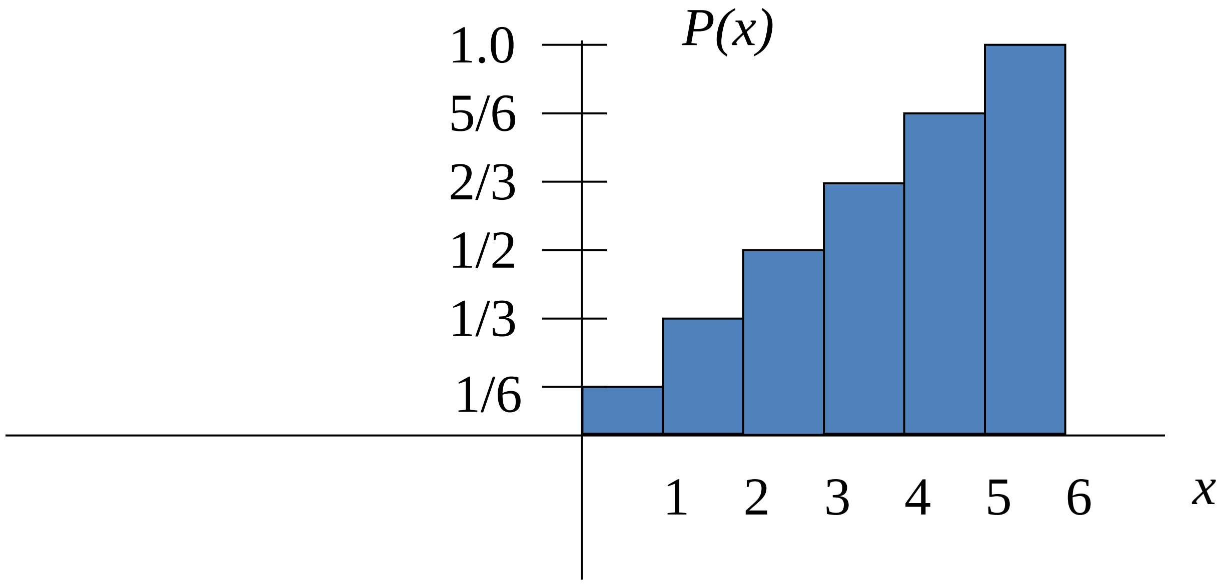
Probability mass function (pmf)

x	$p(x)$
1	$p(x=1)$=1/6
2	$p(x=2)$=1/6
3	$p(x=3)$=1/6
4	$p(x=4)$=1/6
5	$p(x=5)$=1/6
6	$\underline{p(x=6)\text{=1/6}}$

Cumulative distribution function

x	$P(x \leq A)$
1	$P(x \leq 1) = 1/6$
2	$P(x \leq 2) = 2/6$
3	$P(x \leq 3) = 3/6$
4	$P(x \leq 4) = 4/6$
5	$P(x \leq 5) = 5/6$
6	$P(x \leq 6) = 6/6$

Cumulative distribution function (CDF)



Practice Problem:

- The number of patients seen in a clinic in any given hour is a random variable represented by *x*. The probability distribution for *x* is:

x	10	11	12	13	14
$P(x)$	.4	.2	.2	.1	.1

Find the probability that in a given hour:

a. exactly 14 patients arrive $p(x=14) = .1$

b. At least 12 patients arrive $p(x \geq 12) = (.2 + .1 + .1) = .4$

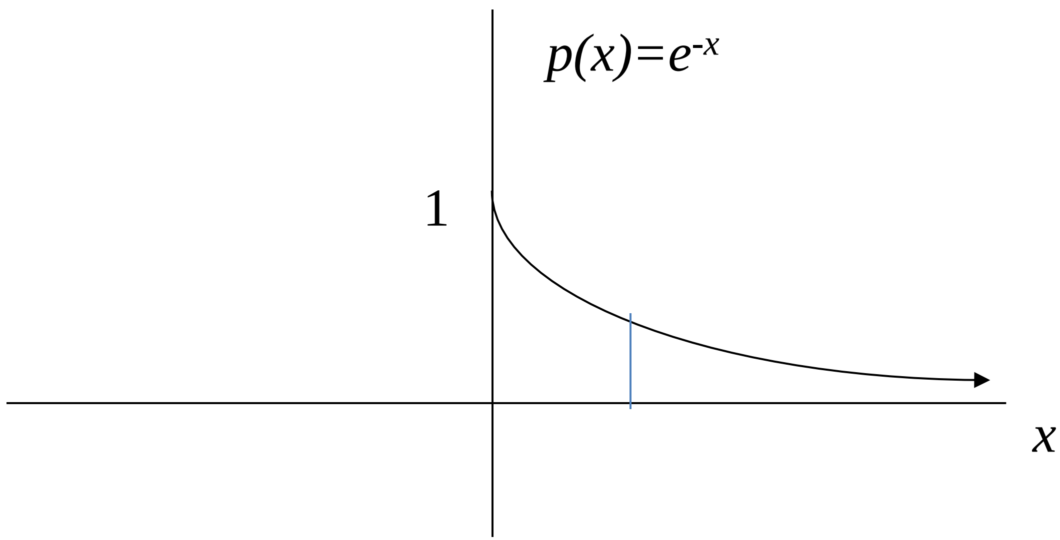
c. At most 11 patients arrive $p(x \leq 11) = (.4 + .2) = .6$

Continuous case

- The probability function that accompanies a continuous random variable is a continuous mathematical function that integrates to 1.
  - For example, recall the negative exponential function (in probability, this is called an "exponential distribution"): $f(x) = e^{-x}$
  - This function integrates to 1:

$$\int_0^{+\infty} e^{-x} = -e^{-x} \Big|_0^{+\infty} = 0 + 1 = 1$$

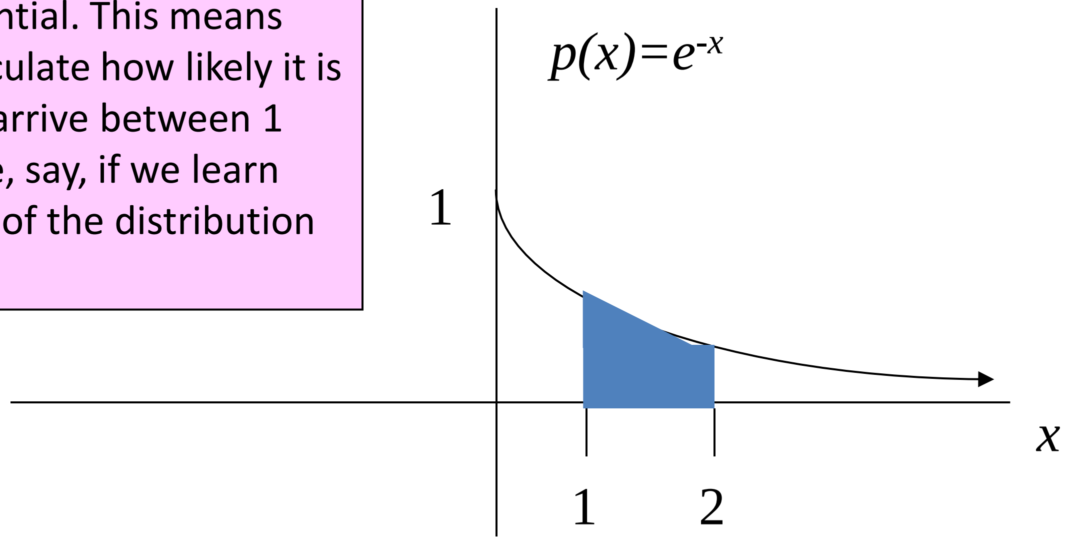
Continuous case: "probability density function" (pdf)



The probability that *x* is any exact particular value (such as 1.9976) is 0; we can only assign probabilities to possible ranges of x.

For example, the probability of *x* falling within 1 to 2:

We saw that train delay times are roughly exponential. This means that we can calculate how likely it is for the train to arrive between 1 and 2 hours late, say, if we learn the parameters of the distribution correctly.

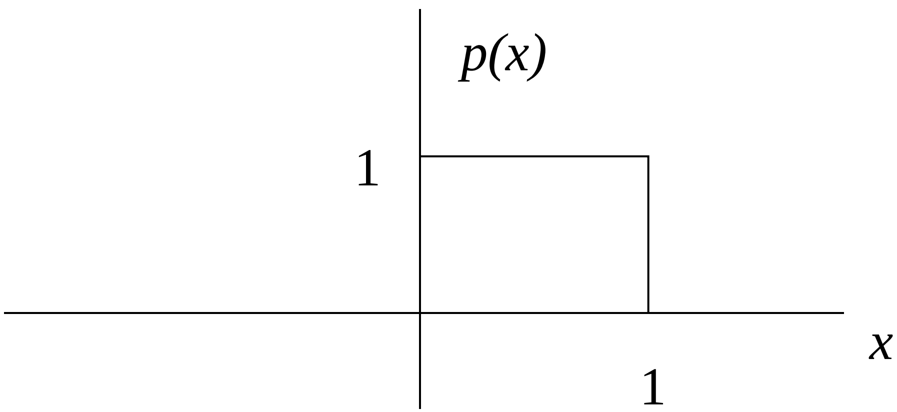


$$P(1 \le x \le 2) = \int_1^2 e^{-x} = -e^{-x} \Big|_1^2 = -e^{-2} - -e^{-1} = -.135 + .368 = .23$$

Example 2: Uniform distribution

The uniform distribution: all values are equally likely.

$f(x) = 1$, for $1 \geq x \geq 0$

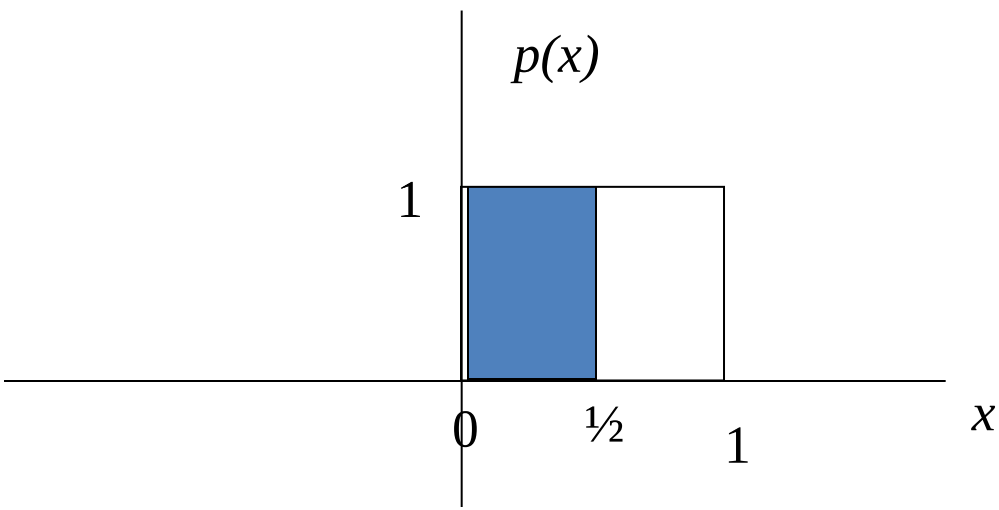


We can see it's a probability distribution because it integrates to 1 (the area under the curve is 1):

$$\int_0^1 1 = x \Big|_0^1 = 1 - 0 = 1$$

Example: Uniform distribution

What's the probability that x is between 0 and ½?



$P(\frac{1}{2} \geq x \geq 0) = \frac{1}{2}$

Expected value

- Recall the following probability distribution of patient arrivals:

x	10	11	12	13	14
$P(x)$	.4	.2	.2	.1	.1

$$\sum_{i=1}^{5} x_i\, p(x) = 10(.4) + 11(.2) + 12(.2) + 13(.1) + 14(.1) = 11.3$$

Example: the lottery

- <u>The Lottery </u>(also known as a tax on people who are bad at math…)
- A certain lottery works by picking 6 numbers from 1 to 49. It costs Rs 1 to play the lottery, and if you win, you win Rs 20 lakhs after taxes.

- *If you play the lottery once, what are your expected winnings or losses?*

Lottery

Calculate the probability of winning in 1 try:

$$\frac{1}{\binom{49}{6}} = \frac{1}{\frac{49!}{43!6!}} = \frac{1}{13{,}983{,}816} = 7.2 \times 10^{-8}$$

"49 choose 6"

Out of 49 numbers, this is the number of distinct combinations of 6.

The probability function (note, sums to 1.0):

Rs x	*p(x)*
-1	.999999928
+ 20 lakh	7.2×10^{-8}

Expected Value

<u>The probability function</u>

x	$p(x)$
-1	.999999928
+ 20 lakh	7.2×10^{-8}

<u>Expected Value</u>

E(X) = P(win)*20,00,000 + P(lose)*-$1.00

$= 2.0 \times 10^{6} * 7.2 \times 10^{-8} + .999999928\ (-1) = .144 - .999999928 = -$ *Rs 0.86*

Negative expected value is never good!

You shouldn't play if you expect to lose money!

Gambling (or how casinos can afford to give so many free drinks…)

A roulette wheel has the numbers 1 through 36, as well as 0 and 00. If you bet Rs 1 that an odd number comes up, you win or lose Rs 1 according to whether or not that event occurs. If random variable X denotes your net gain, X=1 with probability 18/38 and X= -1 with probability 20/38.

$E(X) = 1(18/38) – 1\ (20/38) = -\$.053$

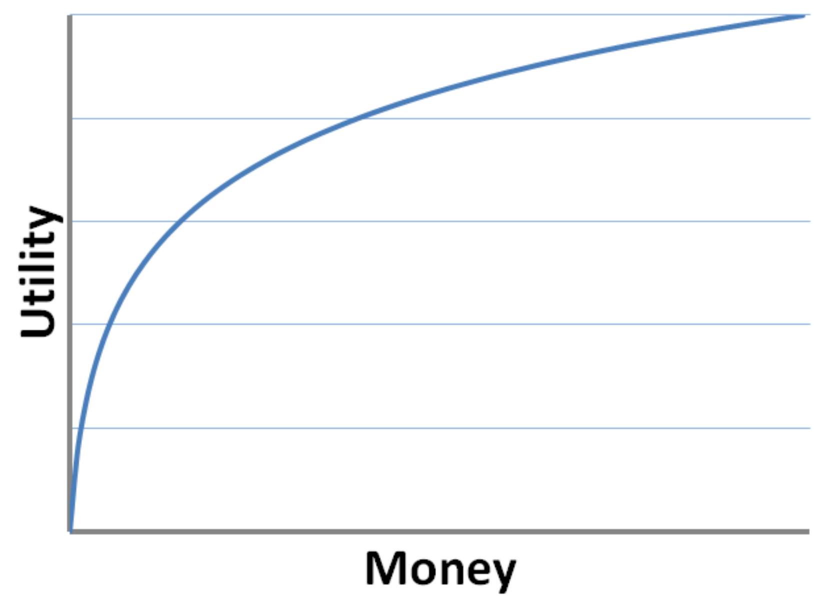
On average, the casino wins (and the player loses) 5 cents per game.

The casino rakes in even more if the stakes are higher:

$E(X) = 10(18/38) – 10\ (20/38) = -\text{Rs } 0.53$

If the cost is Rs 10 per game, the casino wins an average of 53 cents per game. If 10,000 games are played in a night, that's Rs 5300 for simply spinning a colored wheel

Not all powerful



Non-trivial discrete probabilities

Take the example of 5 coin tosses. What's the probability that you flip exactly 3 heads in 5 coin tosses?

A discrete distribution: binomial

- A fixed number of observations (trials), n
  - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- A binary outcome
  - e.g., head or tail in each toss of a coin; disease or no disease
  - Generally called "success" and "failure"
  - Probability of success is p, probability of failure is 1 – p
- Constant probability for each observation
  - e.g., Probability of getting a tail is the same each time we toss the coin

Binomial distribution

Solution:

One way to get exactly 3 heads: HHHTT

What's the probability of this <u>exact</u> arrangement?

P(heads)xP(heads) xP(heads)xP(tails)xP(tails) =$(1/2)^3 \, x \, (1/2)^2$

Another way to get exactly 3 heads: THHHT

Probability of this exact outcome = $(1/2)^1 \, x \, (1/2)^3 \, x \, (1/2)^1 = (1/2)^3 \, x \, (1/2)^2$

Binomial distribution

In fact, $(1/2)^3 \, x \, (1/2)^2$ is the probability of each unique outcome that has exactly 3 heads and 2 tails.

So, the overall probability of 3 heads and 2 tails is:
$(1/2)^3 \, x \, (1/2)^2 + (1/2)^3 \, x \, (1/2)^2 + (1/2)^3 \, x \, (1/2)^2 +$ ….. for as many unique arrangements as there are—but how many are there??

Outcome	Probability
THHHT	$(1/2)^3\ x\ (1/2)^2$
HHHTT	$(1/2)^3\ x\ (1/2)^2$
TTHHH	$(1/2)^3\ x\ (1/2)^2$
HTTHH	$(1/2)^3\ x\ (1/2)^2$
HHTTH	$(1/2)^3\ x\ (1/2)^2$
HTHHT	$(1/2)^3\ x\ (1/2)^2$
THTHH	$(1/2)^3\ x\ (1/2)^2$
HTHTH	$(1/2)^3\ x\ (1/2)^2$
HHTHT	$(1/2)^3\ x\ (1/2)^2$
THHTH	$(1/2)^3\ x\ (1/2)^2$

10 arrangements $x\ (1/2)^3\ x\ (1/2)^2$

$\binom{5}{3}$ ways to arrange 3 heads in 5 trials

The probability of each unique outcome (note: they are all equal)

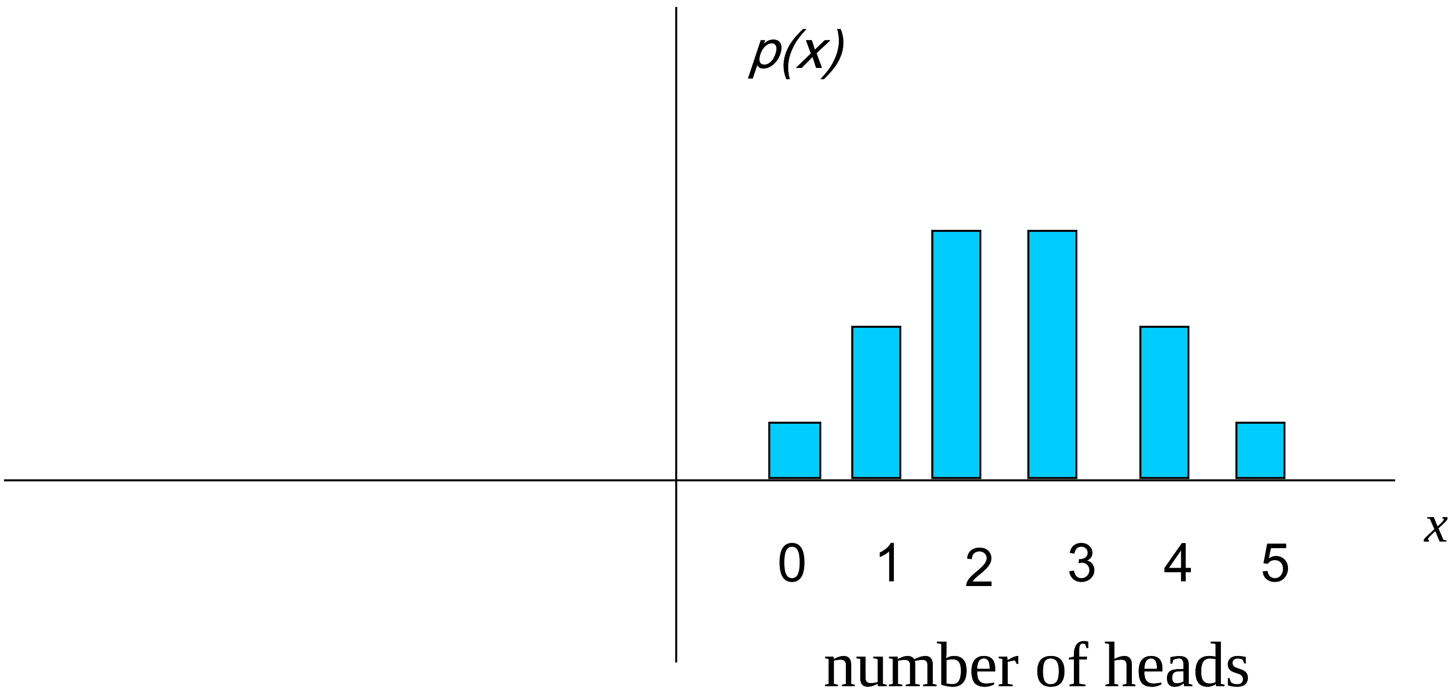
${}_5C_3 = 5!/3!2! = 10$

$$\therefore P(\text{3 heads and 2 tails}) = \binom{5}{3} x\ P(heads)^3\ x\ P(tails)^2 =$$

$$10\ x\ (½)^5 = 31.25\%$$

Binomial distribution function:

X= the number of heads tossed in 5 coin tosses



Binomial distribution, generally

Note the general pattern emerging → if you have only two possible outcomes (call them 1/0 or yes/no or success/failure) in *n* independent trials, then the probability of exactly *X* "successes"=

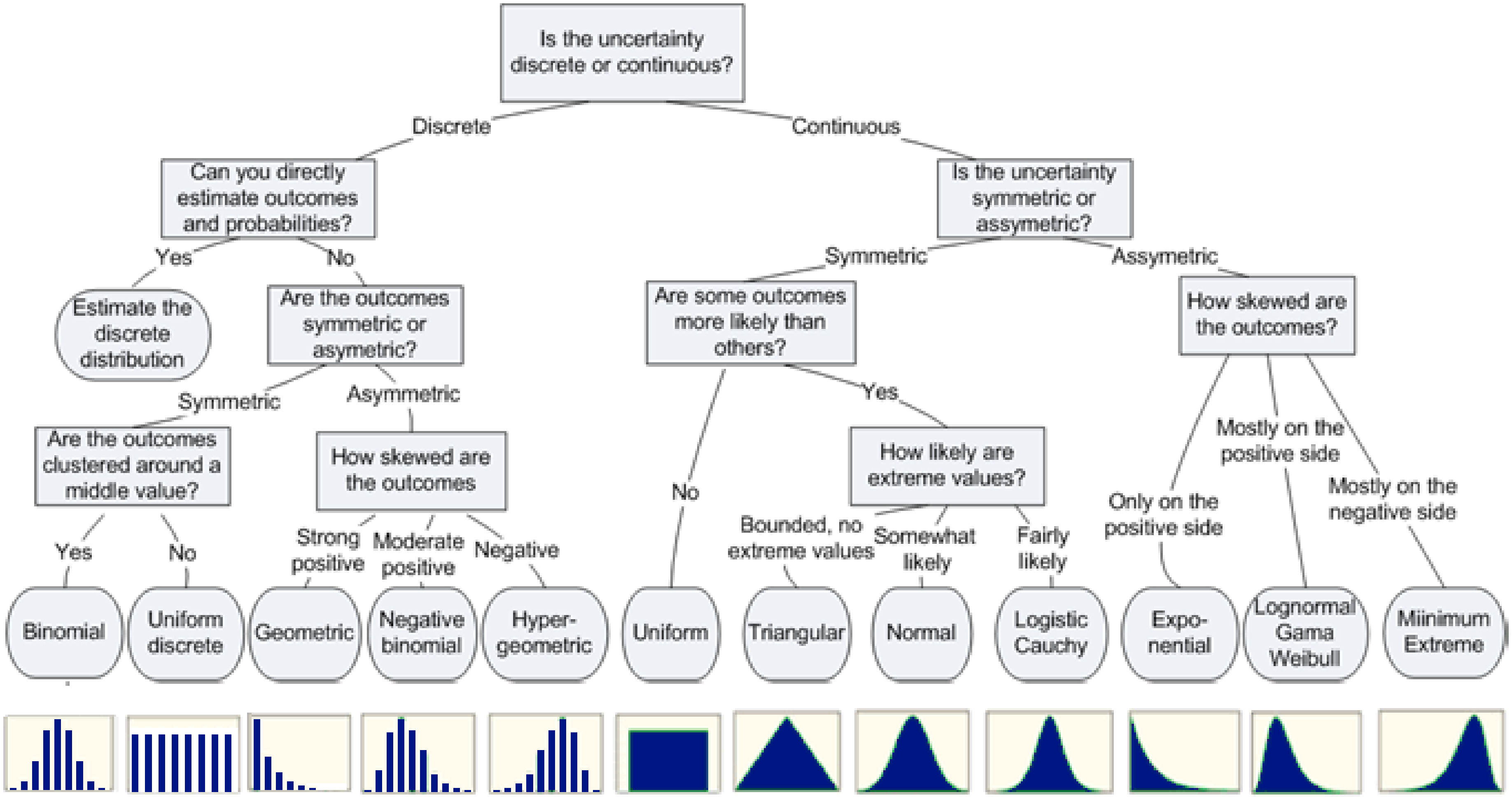
n = number of trials

$$\binom{n}{X} p^X (1-p)^{n-X}$$

1-p = probability of failure

X = # successes out of *n* trials

p = probability of success


Is the uncertainty discrete or continuous?
Discrete
Continuous
Can you directly estimate outcomes and probabilities?
Is the uncertainty symmetric or assymetric?
Yes
No
Symmetric
Assymetric
Estimate the discrete distribution
Are the outcomes symmetric or asymetric?
Are some outcomes more likely than others?
How skewed are the outcomes?
Symmetric
Asymmetric
Are the outcomes clustered around a middle value?
How skewed are the outcomes
No
Yes
How likely are extreme values?
Only on the positive side
Mostly on the positive side
Mostly on the negative side
Yes
No
Strong positive
Moderate positive
Negative
Bounded, no extreme values
Somewhat likely
Fairly likely
Binomial
Uniform discrete
Geometric
Negative binomial
Hyper-geometric
Uniform
Triangular
Normal
Logistic Cauchy
Expo-nential
Lognormal Gama Weibull
Miinimum Extreme

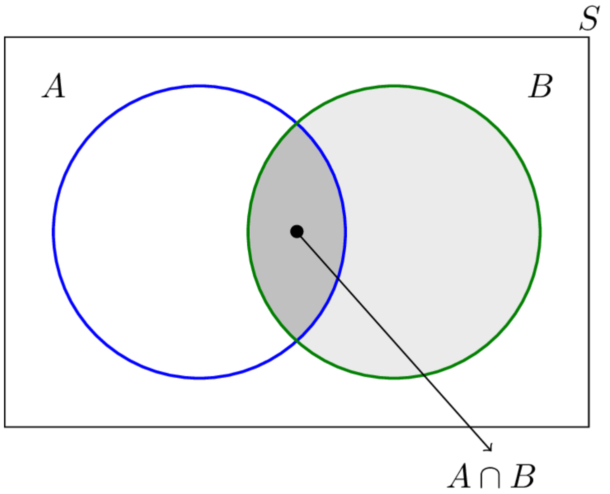
Today's Lecture

- General announcement
  - Final registrations
  - Dropbox file request system
  - Audit requests
- Project announcements
  - Both demos now online
  - Deadlines
    - 20th Jan: tell me what you're doing (1 paragraph; optional)
    - 31st Jan: final submission (code+ 2-3 page summary)
  - Project teams
    - Possible novelty
- Conditional, marginal and joint probabilities
  - How to calculate, how to interpret
  - Derivation of Bayes' theorem
- Bayesian networks
  - Construction and notation
  - Estimation and inference
  - Applications in human-computer interaction
- Reading: Russell & Norvig 14.1 to 14.4
  - Slides from Padhraic Smythe's 2007 talk

Joint probability

	$y=1$	$y=2$	$y=3$
$x=1$	0.30	0.05	0.00
$x=2$	0.05	0.20	0.05
$x=3$	0.00	0.05	0.30

Conditional probability



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

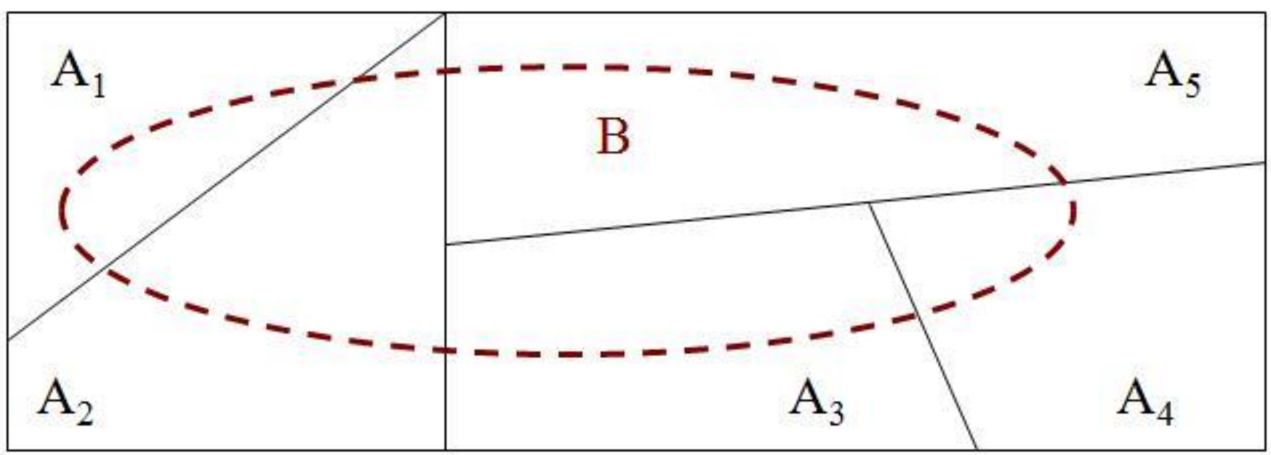
Marginalization

Law of Total Probability

$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|\ A_i)\ P(A_i)$ where

$A_i \cap A_j = \varnothing$ (Mutually Exclusive), and

$\cup A_i = \Omega$ (Collectively Exhaustive)



The joint distribution knows everything

Given a joint distribution (e.g., P(a,b,c,d)) we can obtain any "marginal" probability (e.g., P(b)) by summing out the other variables, e.g.,

$$P(b) = \Sigma_a \Sigma_c \Sigma_d P(a, b, c, d)$$

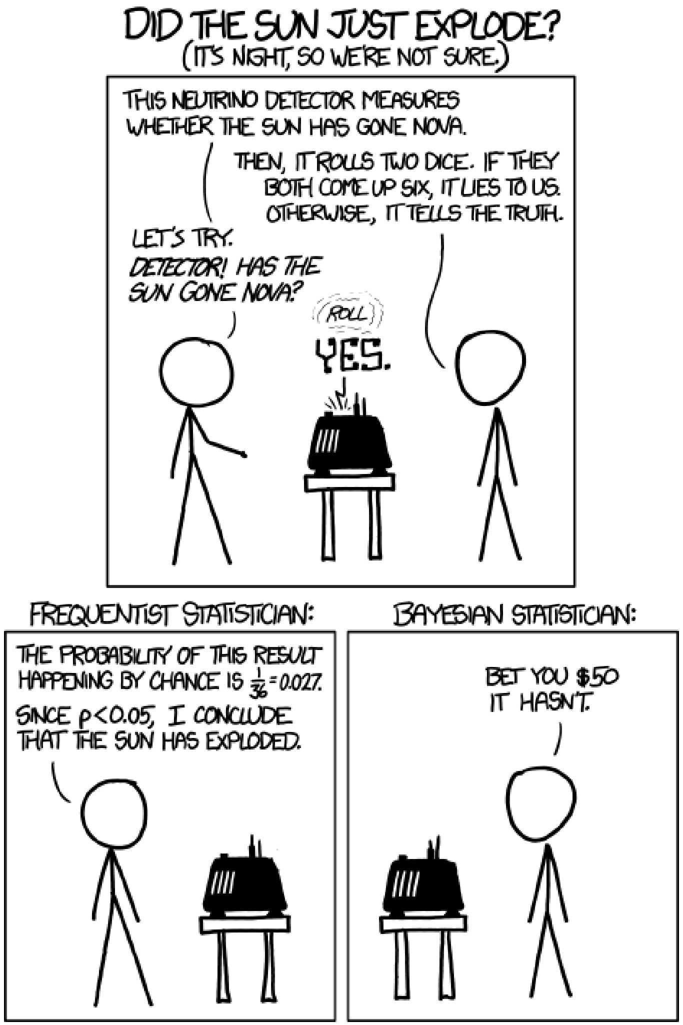
Less obvious: we can also compute <u>any conditional probability of interest</u> given a joint distribution, e.g.,

$$P(c \mid b) = \Sigma_a \Sigma_d P(a, c, d \mid b)$$
$$= 1 / P(b) \ \Sigma_a \Sigma_d P(a, c, d, b)$$

where $1 / P(b)$ is just a normalization constant

The joint distribution contains the information we need to compute any probability of interest.

Corollary: Bayes' theorem



- P(a|b)P(b) = P(b|a)P(a)
- Useful way of appearing wise to your friends

P(extreme event |common trait) =

P(common trait| extreme event) x p(extreme event)/p(common event)

- Prior probabilities can be hard to specify objectively

Computing with Probabilities: The Chain Rule or Factoring

We can always write

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b, c, \ldots z)$$

(by definition of joint probability)

Repeatedly applying this idea, we can write

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b \mid c, .. z)\, P(c \mid .. z) .. P(z)$$

This factorization holds for any ordering of the variables

This is the chain rule for probabilities

Conditional Independence

- 2 random variables A and B are conditionally independent given C iff

  $P(a, b \mid c) = P(a \mid c)\, P(b \mid c)$ for all values a, b, c

- More intuitive (equivalent) conditional formulation
  - A and B are conditionally independent given C iff
    $P(a \mid b, c) = P(a \mid c)$ OR $P(b \mid a, c)$ $P(b \mid c)$, for all values a, b, c

  - Intuitive interpretation:
    $P(a \mid b, c) = P(a \mid c)$ tells us that learning about b, given that we already know c, provides no change in our probability for a,

    i.e., b contains no information about a beyond what c provides

- Can generalize to more than 2 random variables
  - E.g., K different symptom variables X1, X2, … XK, and C = disease
  - $P(X1, X2, \ldots XK \mid C) = \prod P(Xi \mid C)$
  - Also known as the naïve Bayes assumption

Bayesian Networks

- A Bayesian network specifies a joint distribution in a structured form
- Represent dependence/independence via a directed graph
  - Nodes = random variables
  - Edges = direct dependence
- Structure of the graph ⇔ Conditional independence relations

In general,

$$p(X_1, X_2, \ldots X_N) = \prod p(X_i \mid \text{parents}(X_i))$$

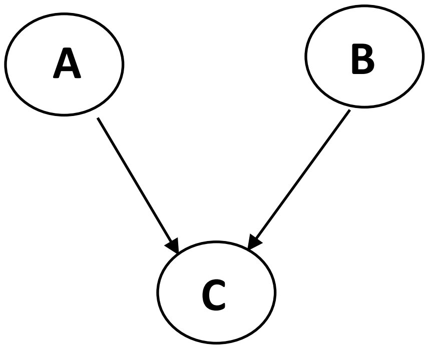
The full joint distribution

The graph-structured approximation

- Requires that graph is acyclic (no directed cycles)
- 2 components to a Bayesian network
  - The graph structure (conditional independence assumptions)
  - The numerical probabilities (for each variable given its parents)

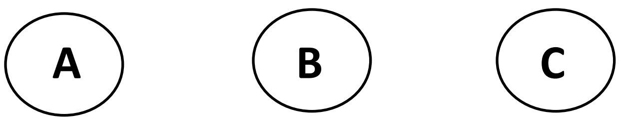
Example of a simple Bayesian network

$p(A,B,C) = p(C|A,B)p(A)p(B)$ $\longleftrightarrow$



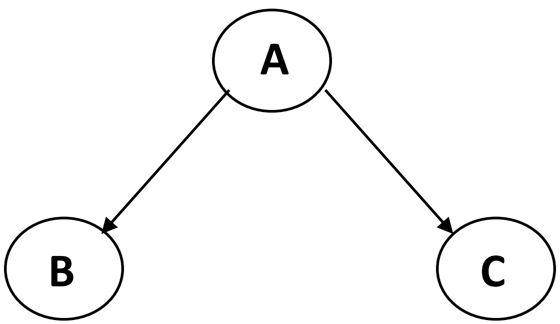
- Probability model has simple factored form
- Directed edges => direct dependence
- Absence of an edge => conditional independence
- Also known as belief networks, graphical models, causal networks
- Other formulations, e.g., undirected graphical models

Examples of 3-way Bayesian Networks



Marginal Independence:
p(A,B,C) = p(A) p(B) p(C)

Examples of 3-way Bayesian Networks

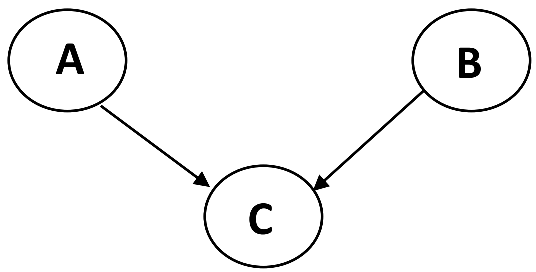


Conditionally independent effects:
p(A,B,C) = p(B|A)p(C|A)p(A)

B and C are conditionally independent Given A

e.g., A is a disease, and we model B and C as conditionally independent symptoms given A

Examples of 3-way Bayesian Networks

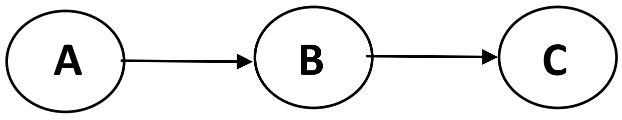


Independent Causes:
p(A,B,C) = p(C|A,B)p(A)p(B)

“Explaining away” effect:
Given C, observing A makes B less likely
e.g., earthquake/burglary/alarm example

A and B are (marginally) independent
but become dependent once C is known

Examples of 3-way Bayesian Networks



Markov dependence:
$p(A,B,C) = p(C|B)\ p(B|A)p(A)$

Example

- Consider the following 5 binary variables:
  - B = a burglary occurs at your house
  - E = an earthquake occurs at your house
  - A = the alarm goes off
  - J = John calls to report the alarm
  - M = Mary calls to report the alarm

  - What is P(B | M, J) ? (for example)

  - We can use the full joint distribution to answer this question
    - Requires $2^5 = 32$ probabilities

    - Can we use prior domain knowledge to come up with a Bayesian network that requires fewer probabilities?

Construct a Bayesian Network: Step 1

- Order the variables in terms of causality
  e.g., {E, B} -> {A} -> {J, M}

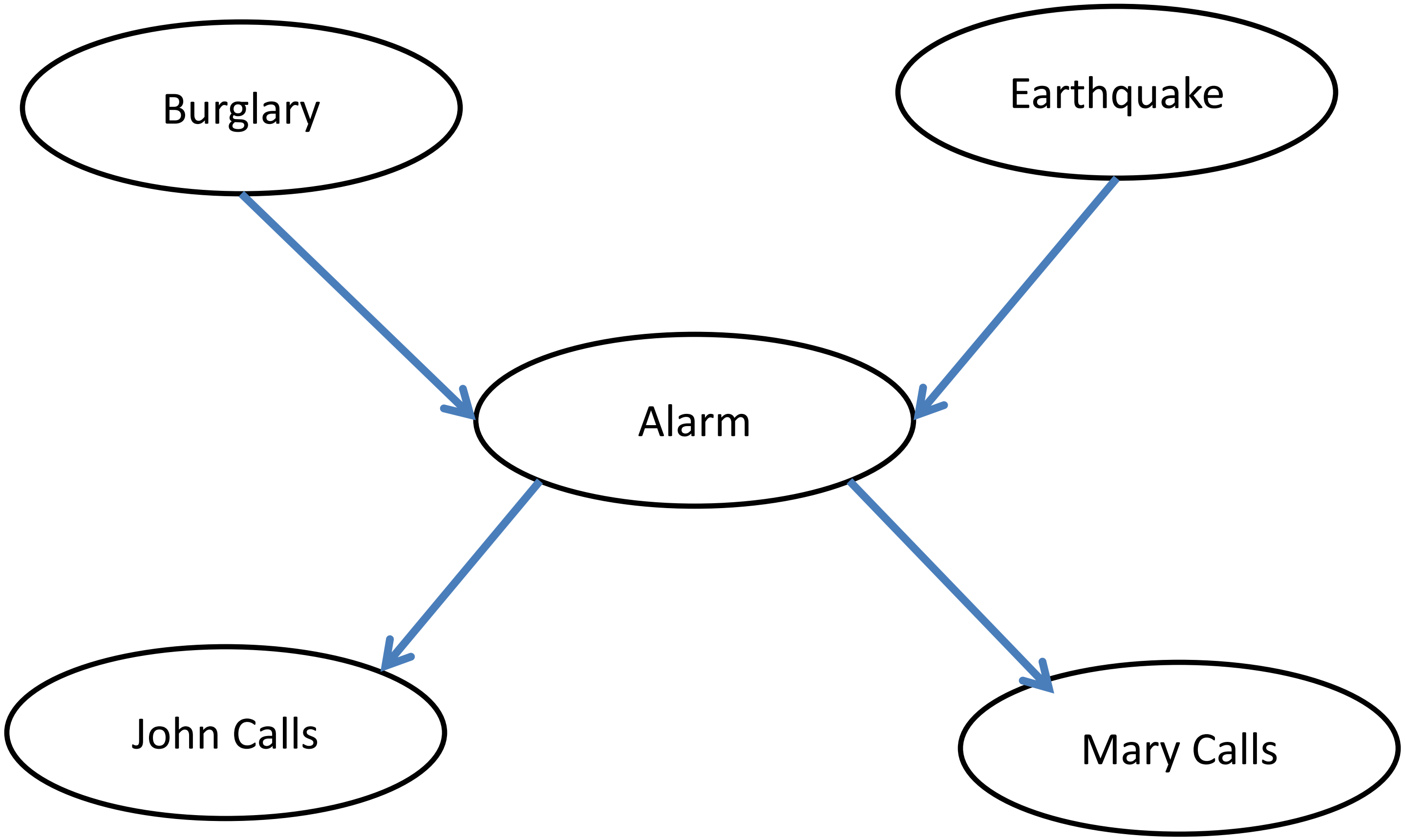
- P(J, M, A, E, B) = P(J, M | A, E, B) P(A| E, B) P(E, B)

  ~ P(J, M | A) P(A| E, B) P(E) P(B)

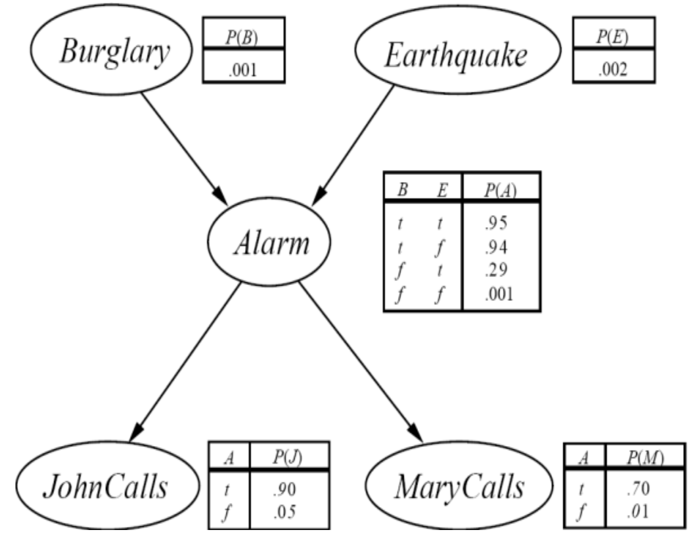
  ~ P(J | A) P(M | A) P(A| E, B) P(E) P(B)

These CI assumptions are reflected in the graph structure of the Bayesian network

Graph structure of network



Constructing this Bayesian Network: Step 2



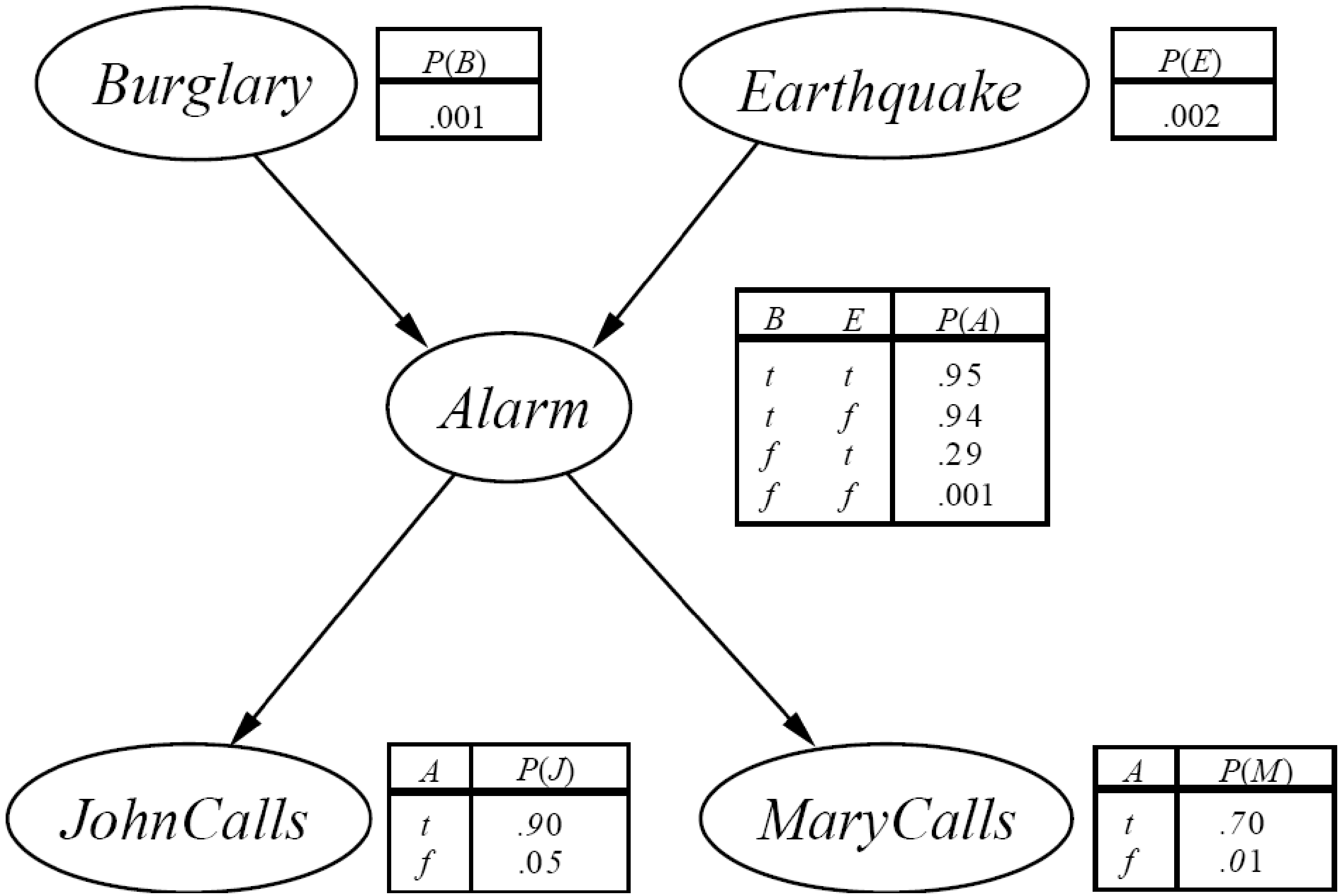
- P(J, M, A, E, B) =
  P(J | A) P(M | A) P(A | E, B) P(E) P(B)

- There are 3 conditional probability tables to be determined:
  P(J | A), P(M | A), P(A | E, B)
  - Requiring 2 + 2 + 4 = 8 probabilities

- And 2 marginal probabilities P(E), P(B) -> 2 more probabilities

- Where do these probabilities come from?
  - Expert knowledge
  - From data (relative frequency estimates or regression analyses)

The Bayesian network



$P(B)$
.001

$P(E)$
.002

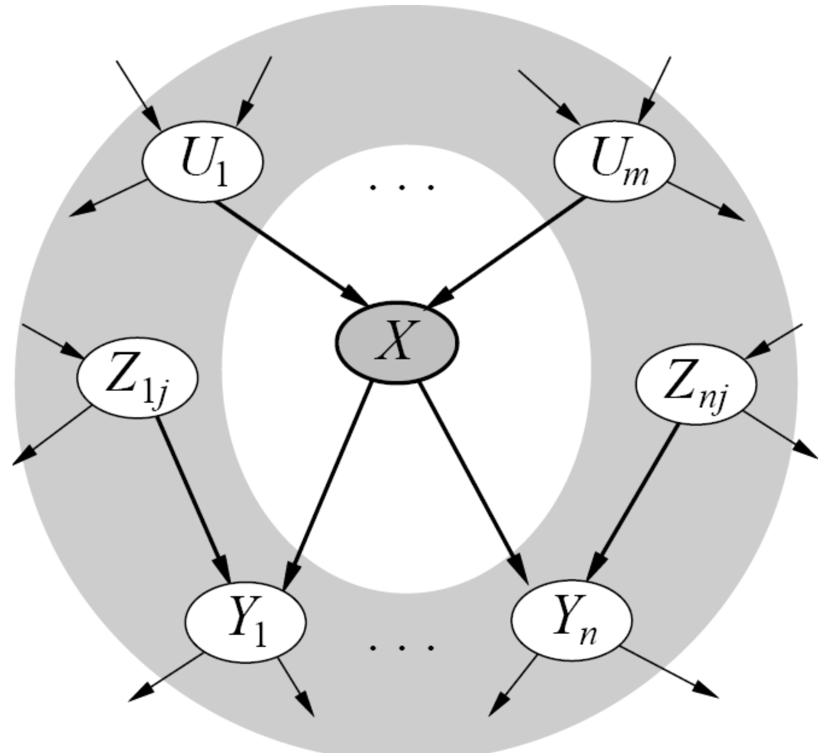
B	E	$P(A)$
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	$P(J)$
t	.90
f	.05

A	$P(M)$
t	.70
f	.01

Intuitive display of conditional independence

A node is conditionally independent of all other nodes in the network given its Markov blanket (in gray)



Number of probabilities in Bayesian Networks

- Consider n binary variables

- Unconstrained joint distribution requires $O(2^n)$ probabilities

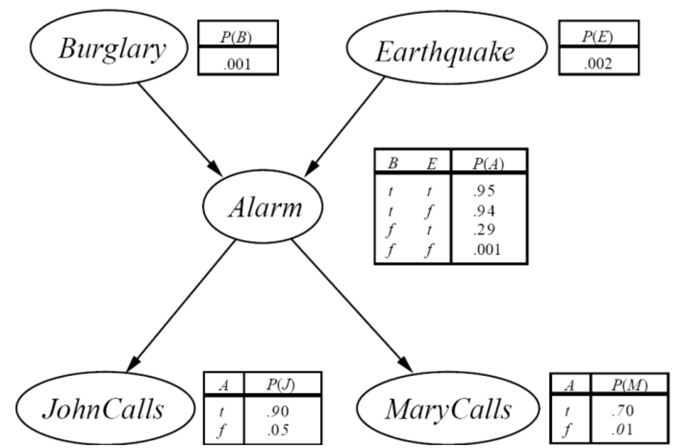
- If we have a Bayesian network, with a maximum of k parents for any node, then we need $O(n\ 2^k)$ probabilities

- Example
  - Full unconstrained joint distribution
    - n = 30: need 10^9 probabilities for full joint distribution
  - Bayesian network
    - n = 30, k = 4: need 480 probabilities

Inference (Reasoning) in Bayesian Networks

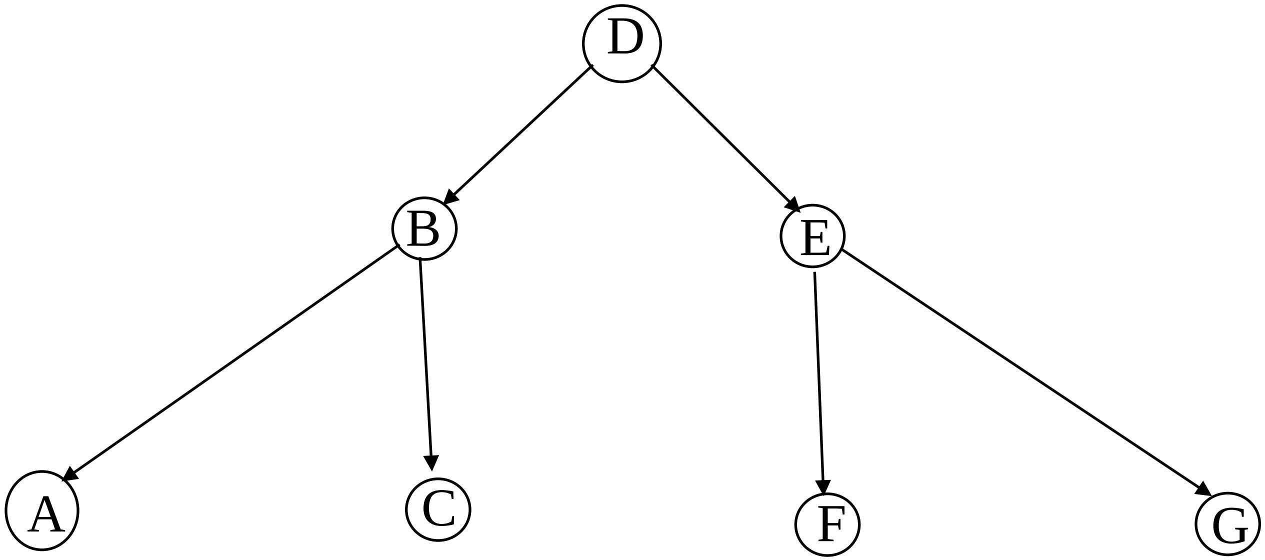
- Consider answering a query in a Bayesian Network
  - Q = set of query variables
  - e = evidence (set of instantiated variable-value pairs)
  - Inference = computation of conditional distribution P(Q | e)

- Examples
  - P(burglary | alarm)
  - P(earthquake | JCalls, MCalls)
  - P(JCalls, MCalls | burglary, earthquake)



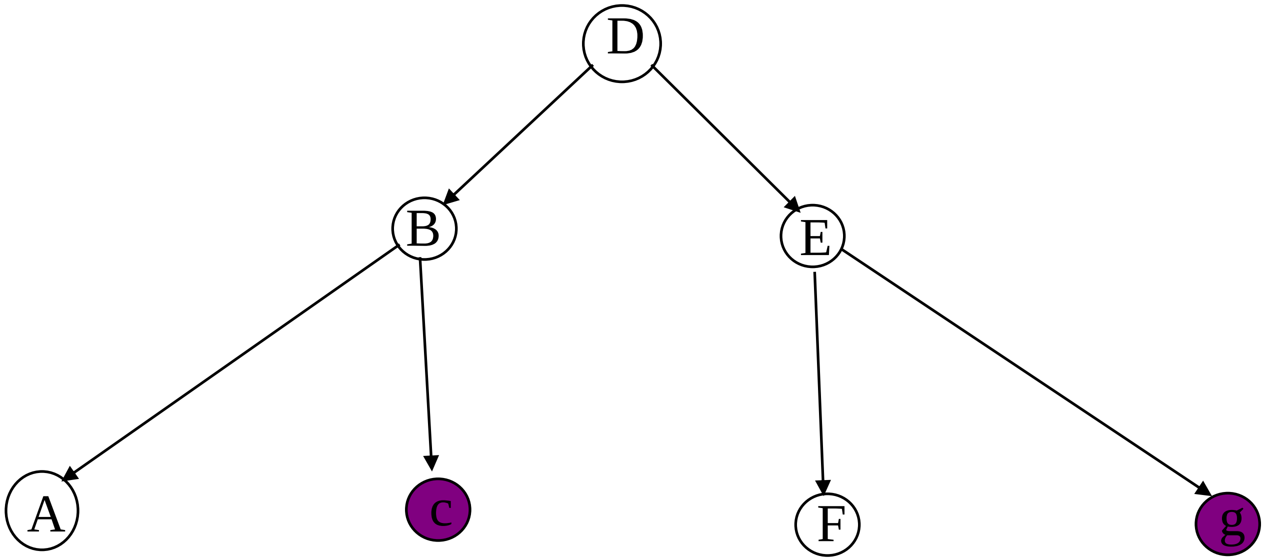
- Can we use the structure of the Bayesian Network to answer such queries efficiently? Answer = yes
  - Generally speaking, complexity is inversely proportional to sparsity of graph

Example: Tree-Structured Bayesian Network



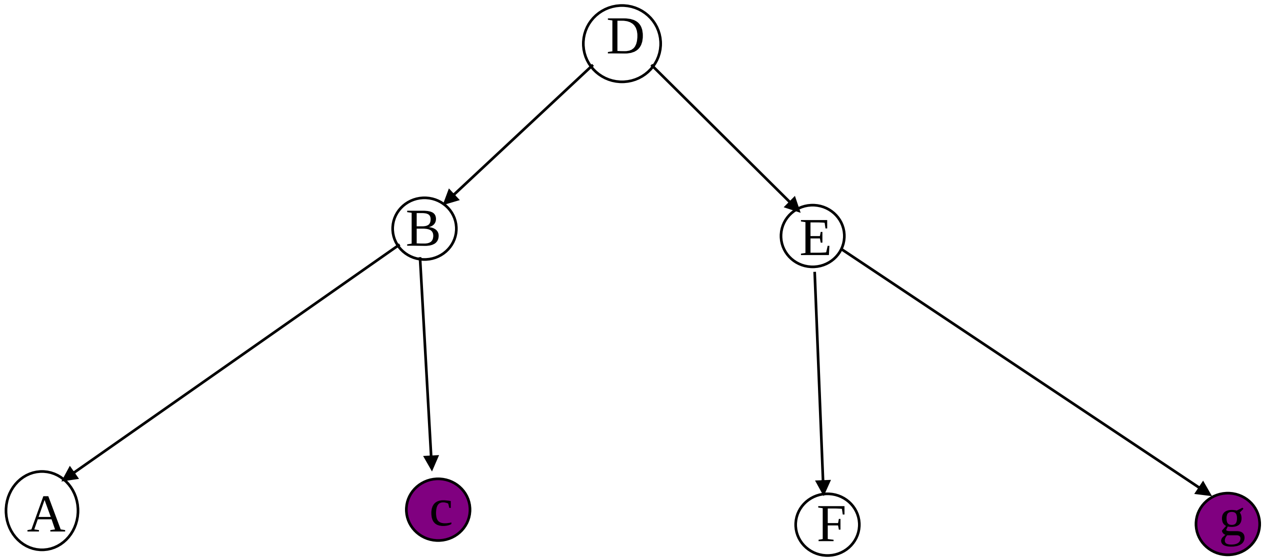
p(a, b, c, d, e, f, g) is modeled as p(a|b)p(c|b)p(f|e)p(g|e)p(b|d)p(e|d)p(d)

Example



Say we want to compute p(a | c, g)

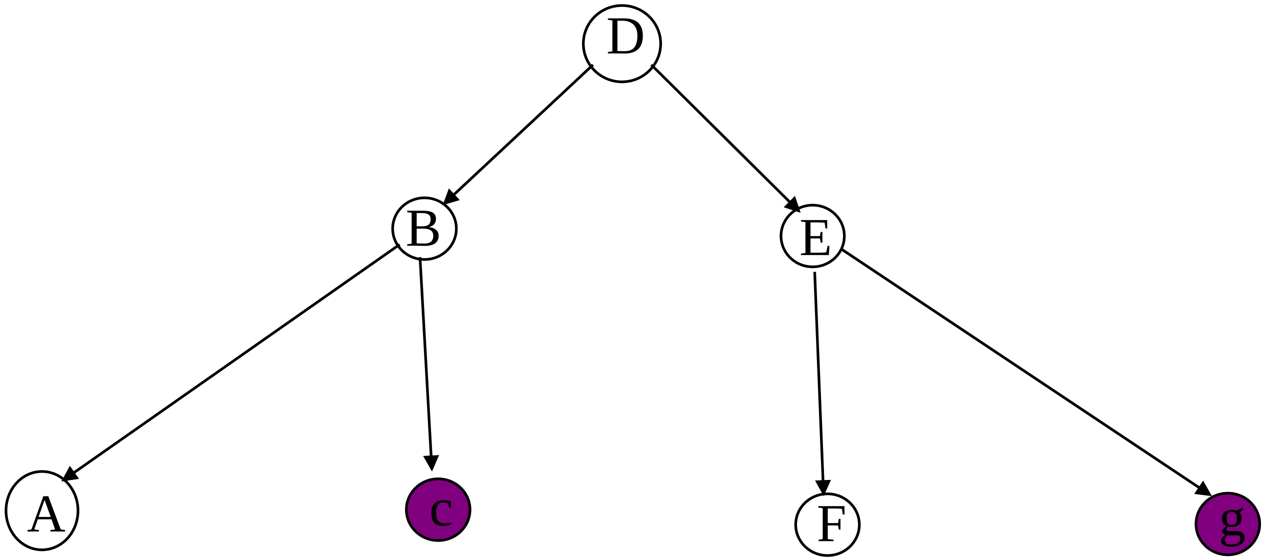
Example



Direct calculation: $p(a|c,g) = \Sigma_{bdef}\, p(a,b,d,e,f \mid c,g)$

Complexity of the sum is $O(m^4)$

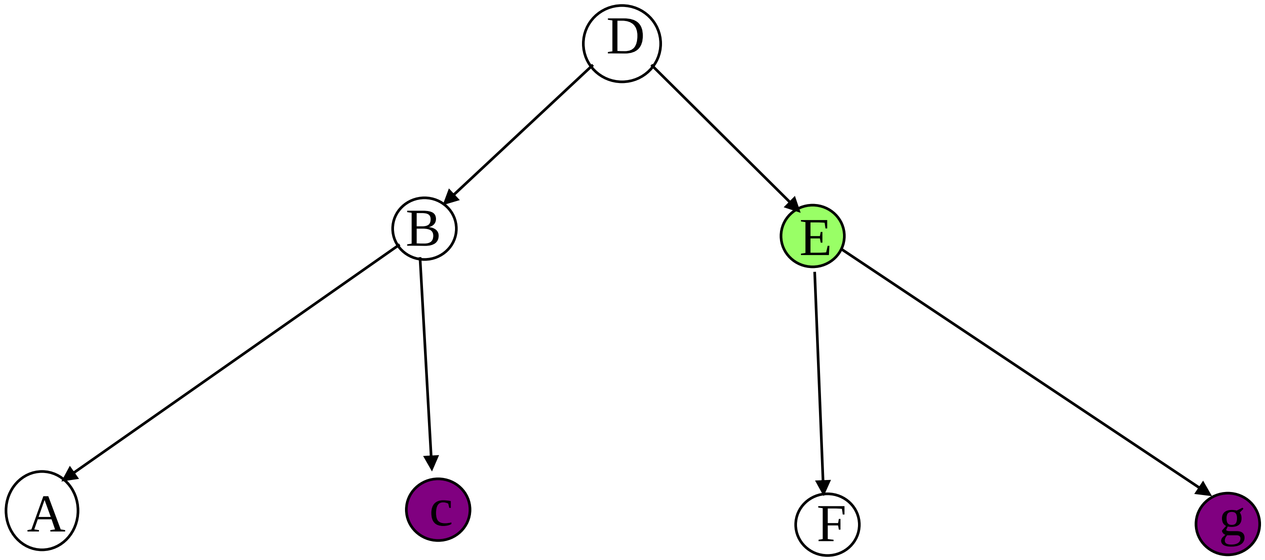
Example



Reordering:

$$\Sigma_b\, p(a|b)\, \Sigma_d\, p(b|d,c)\, \Sigma_e\, p(d|e)\, \Sigma_f\, p(e|f,g)p(f|g)$$

Example

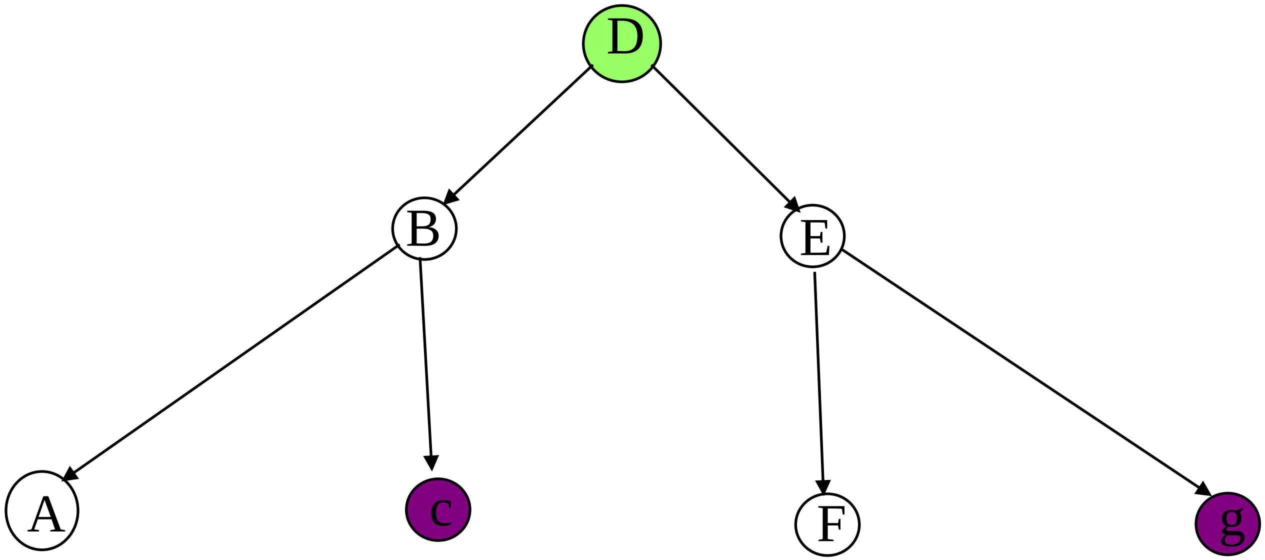


Reordering:

$$\Sigma_b\ p(a|b)\ \Sigma_d\ p(b|d,c)\ \Sigma_e\ p(d|e)\ \Sigma_f\ p(e,f\ |g)$$

p(e|g)

Example

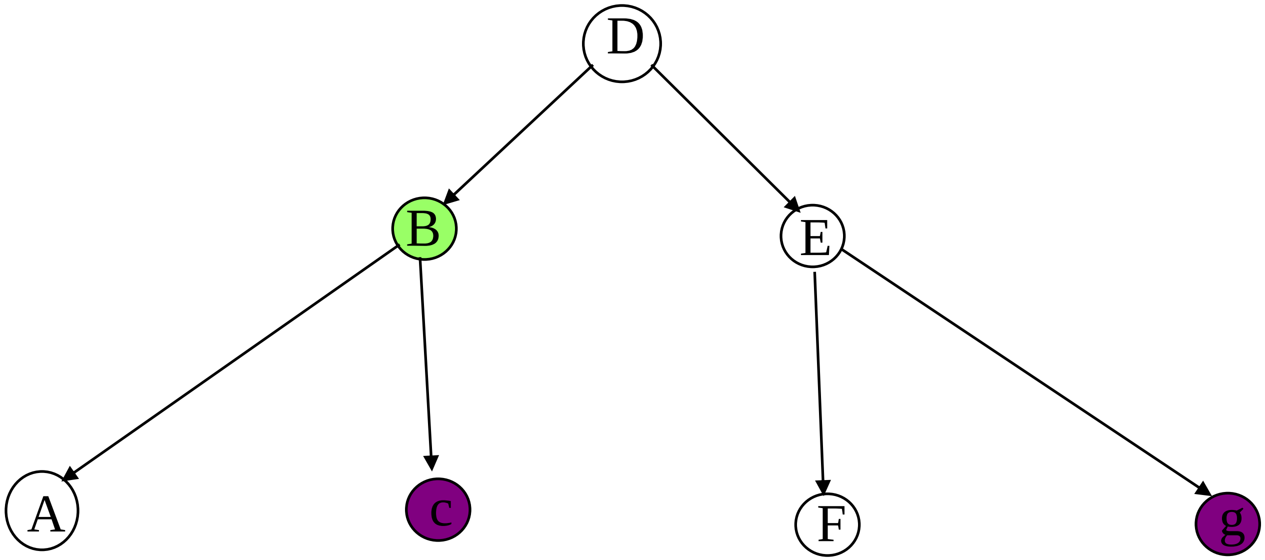


Reordering:

$$\Sigma_b\ p(a|b)\ \Sigma_d\ p(b|d,c)\ \Sigma_e\ p(d|e)\ p(e|g)$$

p(d|g)

Example

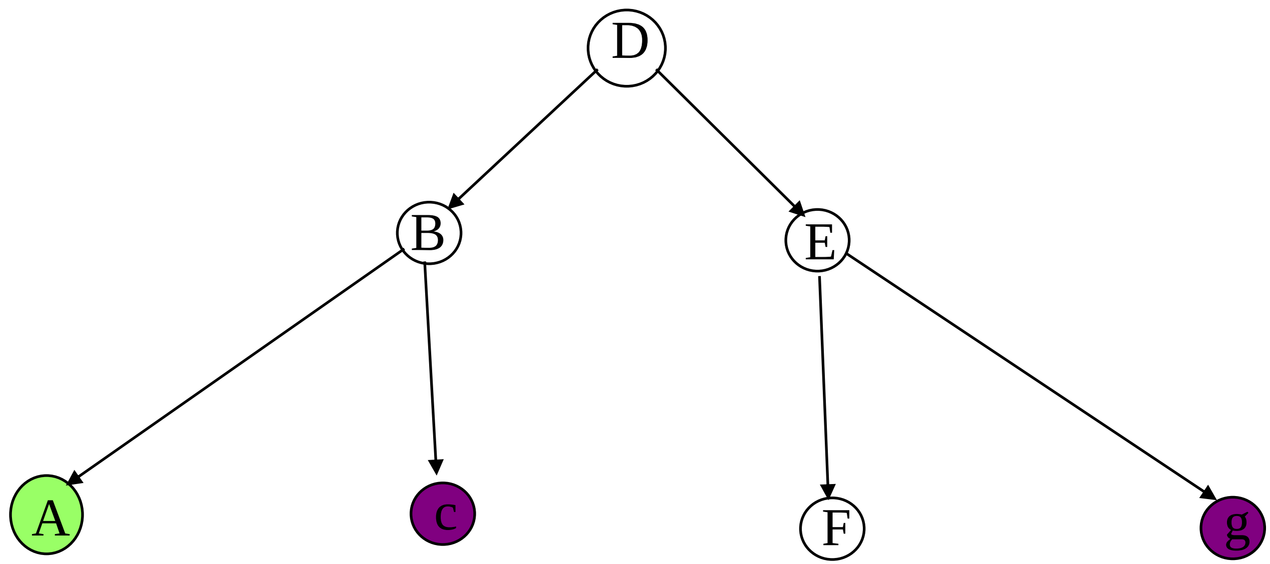


Reordering:

$$\Sigma_b \; p(a|b) \; \Sigma_d \; p(b|d,c) \; p(d|g)$$

$p(b|c,g)$

Example



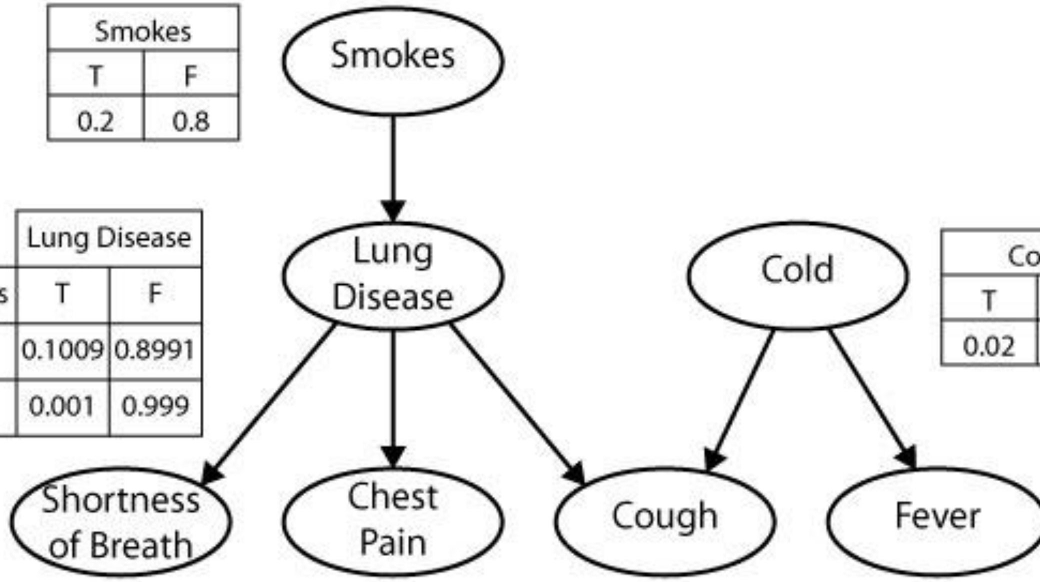
Reordering:

$\sum_b p(a|b)\, p(b|c,g)$

$p(a|c,g)$

Complexity is $O(m)$, compared to $O(m^4)$

Example with numbers



Smokes	
T	F
0.2	0.8

	Lung Disease	
Smokes	T	F
T	0.1009	0.8991
F	0.001	0.999

Cold	
T	F
0.02	0.98

	Shortness of Breath	
Lung Disease	T	F
T	0.208	0.792
F	0.01	0.99

	Chest Pain	
Lung Disease	T	F
T	0.208	0.792
F	0.01	0.99

		Cough	
Lung Disease	Cold	T	F
T	T	0.7525	0.2475
T	F	0.505	0.495
F	T	0.505	0.495
F	F	0.01	0.99

	Fever	
Cold	T	F
T	0.307	0.693
F	0.01	0.99

General Strategy for inference

- Want to compute P(q | e)

Step 1:

$P(q \mid e) = P(q,e)/P(e) = \alpha\, P(q,e)$, since P(e) is constant wrt Q

Step 2:

$P(q,e) = \Sigma_{a..z}\ P(q, e, a, b, \ldots. z)$, by the law of total probability

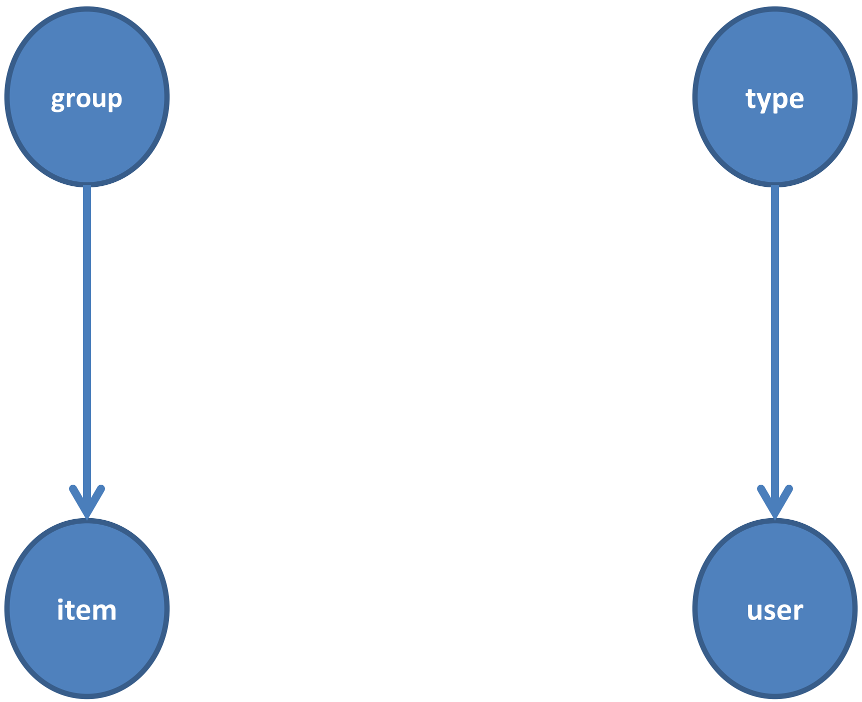
Step 3:

$\Sigma_{a..z}\ P(q, e, a, b, \ldots. z) = \Sigma_{a..z}\ \Pi_i\ P(\text{variable } i \mid \text{parents } i)$

(using Bayesian network factoring)

Step 4: Distribute summations across product terms for efficient computation

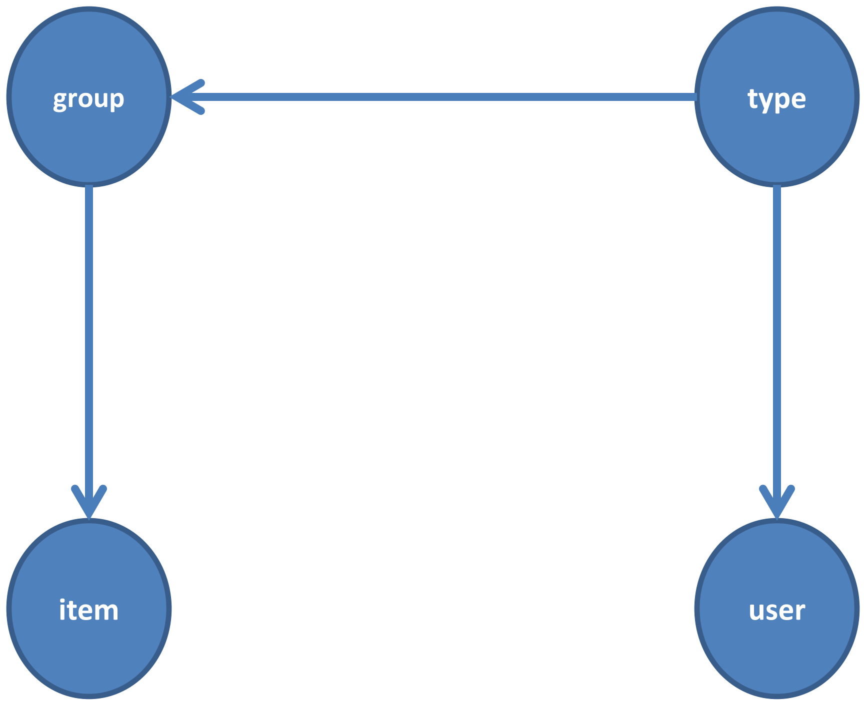
Recommender system example



Can you calculate p(item/user) for a particular user?

$$p(i|u) = \frac{1}{p(u)} \sum_{g,t} p(i|g)p(u|t)p(g)p(t)$$

Recommender system example



Try now

$$p(i|u) = \frac{1}{p(u)} \sum_{g,t} p(i|g)\overbrace{p(u|t)}^{\text{Personalization}}\underbrace{p(g|t)}_{\text{Context inference}}p(t)$$