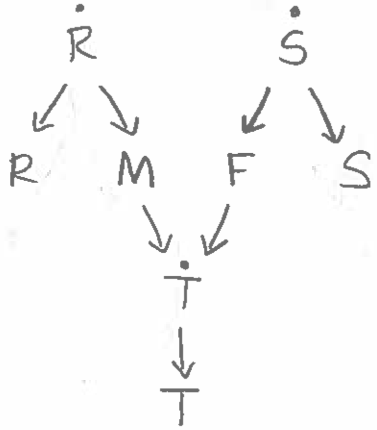
INFERENCE BY SAMPLING

① Let's consider another method for computing conditional probabilities from the "blood type" network. Recall:



where:

- $\dot{R}, \dot{S}, \dot{T}$ are the blood genotypes for Rhonda, Sam, and Tim ($\in \{AA, AB, AO, BB, BO, OO\}$)
- R, S, T are the blood types for Rhonda, Sam, and Tim ($\in \{A, B, AB, O\}$)
- M, F are the genes passed to Tim by Rhonda (his Mother) and Sam (his Father). $M, F \in \{A, B, O\}$.

Let's try to compute $P(R = A \mid T = AB)$, the probability that Rhonda has blood type A given that Tim has blood type AB.

INFERENCE BY SAMPLING

(2) It is easy to sample from the joint distribution $P(\dot{R},\dot{S},\dot{T},R,S,T,M,F)$ by exploiting the factorization provided by the Bayesian network:

$$P(\dot{r},\dot{s},\dot{t},r,s,t,m,f)$$

$$= \underbrace{P(\dot{r})}_{\substack{\text{sample}\\ \dot{r}}}\underbrace{P(\dot{s})}_{\substack{\text{sample}\\ \dot{s}}}\underbrace{P(r|\dot{r})}_{\substack{\text{sample}\\ r,\text{ given}\\ \text{our sample}\\ \dot{r}}}\underbrace{P(m|\dot{r})}_{\substack{\text{sample}\\ m,\text{ given}\\ \text{our sample}\\ \dot{r}}}P(f|\dot{r})P(\dot{t}|m,f)P(t|\dot{t})$$

... etc.

(3) In general, we can sample from the joint distribution encoded by a Bayesian network as follows:

- take a topological order $X_1, \ldots, X_n$ of the variables in the Bayesian Network
- for i in 1 to n:
  - let x_i be a sample from $P(x_i|pa_i)$, where pa_i is the values that the parents of X_i have been set to.

a topological order of a directed acyclic graph is an ordering of the nodes so that a parent never appears after any of its children

INFERENCE BY SAMPLING

④ If we sample from a distribution, we can estimate it!

```
• Initialize $count(x_1, \dots, x_n) = 0 \quad \forall x_1, \dots, x_n \in X_1 \times \dots \times X_n$
• Repeat K times:
    • sample $x_1, \dots, x_n$ from $P(X_1, \dots, X_n)$
    • $count(x_1, \dots, x_n)$ += 1
```

As $K \to \infty$, our estimate $\dfrac{count(x_1, \dots, x_n)}{K} \to P(x_1, \dots, x_n)$

⑤ So if we want to estimate a conditional probability like $P(R=A \mid T=AB)$, we can reexpress it:

$$P(R=A \mid T=AB) = \frac{P(R=A, T=AB)}{P(T=AB)}$$

$$= \frac{\sum_{\dot r, \dot s, \dot t, s, m, f} P(\dot r, \dot s, \dot t, R=A, s, T=AB, m, f)}{\sum_{\dot r, \dot s, \dot t, r, s, m, f} P(\dot r, \dot s, \dot t, r, s, T=AB, m, f)}$$

$$\approx \frac{\sum_{\dot r, \dot s, \dot t, s, m, f} count(\dot r, \dot s, \dot t, R=A, s, T=AB, m, f)}{\sum_{\dot r, \dot s, \dot t, r, s, m, f} count(\dot r, \dot s, \dot t, r, s, T=AB, m, f)}$$

$$= \frac{\text{total samples where } R=A \text{ and } T=AB}{\text{total samples where } T=AB}$$

INFERENCE BY SAMPLING

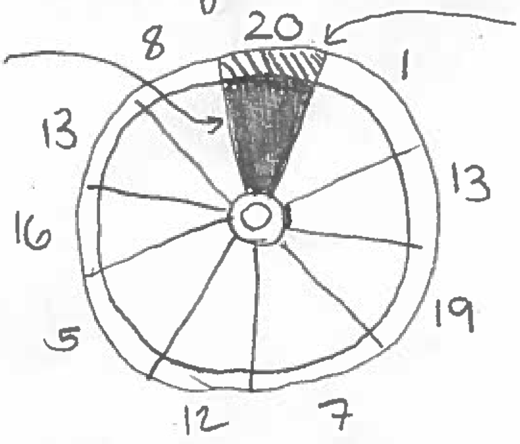
⑥ This inference technique is called <u>rejection sampling</u>:

```
RejectionSampling(joint distribution P, y, x):
  - let N = 0, D = 0
  - repeat K times:
      - sample S from joint distribution P:
          - if X = x and Y = y in S:
                N += 1
          - if X = x in S:
                D += 1
  - return N/D
```

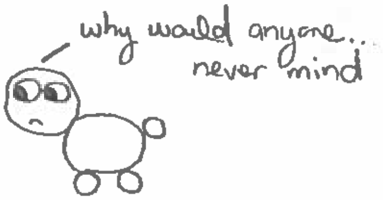
As $K \to \infty$, $\text{RejectionSampling}(P, x, y) \to P(y|x)$

⑦ Rejection sampling is analogous to the following scenario: Imagine somebody shows you a dartboard:

this shaded wedge is the 20-wedge



this striped part is "double-20"



They ask:

The area of "double-20" is what percentage of the area of the 20 wedge?

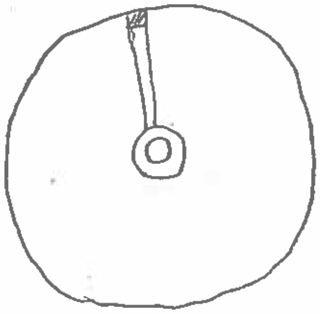
INFERENCE BY SAMPLING

8) Geometry is a pain. Luckily, you have the uncanny ability to throw darts uniformly at random (some may call this a handicap, but don't listen to them).

To answer the question, you can start throwing darts, then:

- count how many darts D hit the 20 wedge
- count how many darts N hit double-20 (which is contained inside the 20 wedge)
- report $\frac{N}{D}$ as the answer

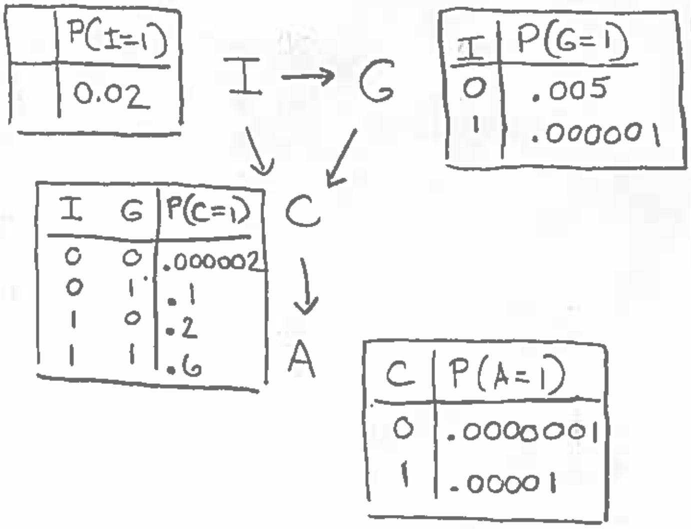
9) But what if the wedge is really thin?



Most of your darts miss, so most of your effort goes to waste.

INFERENCE BY SAMPLING

(10) Consider the following scenario. Infowars (a far right website) and Goop (a lifestyle magazine founded by Gwyneth Paltrow) don't have much in common, but they both advocate the use of colloidal silver (a mixture containing silver particles) for health benefits. One rare but unfortunate side effect of colloidal silver is argyria, a condition in which the skin turns a deep blue. Let's model this with the following network:



	P(I=1)
	0.02

I	P(G=1)
0	.005
1	.000001

I	G	P(C=1)
0	0	.000002
0	1	.1
1	0	.2
1	1	.6

C	P(A=1)
0	.0000001
1	.00001

where:

$I=1$ if the patient subscribes to Infowars

$G=1$ if the patient subscribes to Goop

$C=1$ if the patient uses colloidal silver

$A=1$ if the patient has argyria

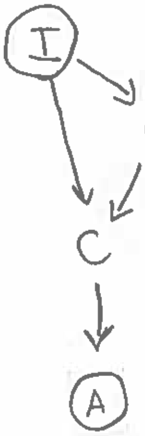
INFERENCE BY SAMPLING

(11) What if we use rejection sampling to estimate $P(I=1 \mid A=1)$, i.e. the probability that an argyria patient subscribes to Infowars?

Well, we only can use samples where $A=1$, since our estimate is $\dfrac{\text{num samples where } I=1 \text{ and } A=1}{\text{num samples where } A=1}$.

That's a pretty rare occurrence: they come along once every million or so (or more) samples. So we waste a lot of samples, and our estimate will take a really long time to converge.

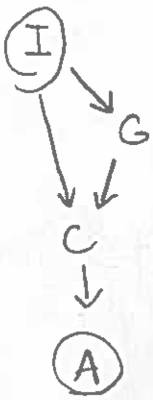
(12) Maybe we can speed things along by "forcing" the values of our evidence and query variables to be what we're interested in. i.e., to estimate $P(I=1, A=1)$:



set $I=1$ (b/c it's an evidence var)	$I=1$
sample g from $P(g \mid I=1)$:	$G=0$
sample c from $P(c \mid I=1, g)$:	$C=0$
set $A=1$ (b/c it's an evidence var)	$A=1$

13 So this seems problematic. We set the evidence variable A without any regard to the previously sampled values. So what is this 'pseudosample good for?

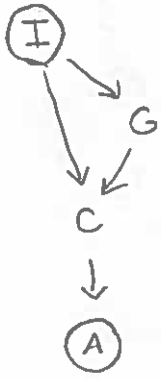
What happens if we weight the sample using the probability of the evidence variables given the previously sampled values?



set	sample	weight
$I=1$		$P(I=1)=.02$
	$G=0$	1
	$C=0$	1
$A=1$		$P(A=1 \mid C=0)=.0000001$

Now we have a sample $(I=1, G=0, C=0, A=1)$ with weight $.02 \cdot 1 \cdot 1 \cdot .0000001 = .00000002$

14 We can keep sampling like this:



set	sample	weight
$I=1$		$P(I=1)=.02$
	$G=0$	1
	$C=1$	1
$A=1$		$P(A=1 \mid C=1)=.0001$

Now we have another sample $(I=1, G=0, C=1, A=1)$ with weight .000002

INFERENCE BY SAMPLING

(15) Proposition:

sampled values of g, c in the nth sample

$$P(I=1, A=1) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} w(I=1, \overset{\downarrow}{g_n}, \overset{\downarrow}{c_n}, A=1)$$

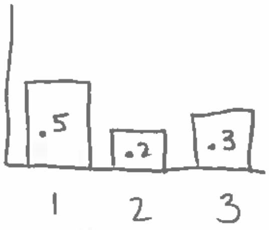
$\uparrow$

weight of the sample

(16) To show this, we'll use a result from statistics. Let $f(x)$ be a function of a variable X with domain $D(X)$. If we draw N independent samples $x_1, \ldots, x_N$ from distribution $P(x)$, then:

$$\lim_{N\to\infty} \sum_{n=1}^{N} f(x_n) = N \cdot \sum_{x\in D(X)} P(x) f(x)$$

e.g. say $X \in \{1,2,3\}$, $f(x) = x$, and $P(x) =$



we sample 10 times:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	3	1	1	2	3	3	1	2	1

$$\sum_{n=1}^{N} f(x_n) = 1+3+1+1+2+3+3+1+2+1 = 18$$

$$N \cdot \sum_{x\in D(x)} P(x) f(x) = 10 \cdot \left(0.5\cdot 1 + 0.2\cdot 2 + 0.3\cdot 3 \right)$$

$$= 10 \cdot (.5 + .4 + .9) = 18$$

(17) Thus:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \omega(I=1, g_n, c_n, A=1)$$

$$= \lim_{N\to\infty} \frac{1}{N} \cdot \left(\lim_{N\to\infty} \sum_{n=1}^{N} \omega(I=1, g_n, c_n, A=1) \right)$$

$$= \lim_{N\to\infty} \frac{1}{\cancel{N}} \cdot \left(\cancel{N} \cdot \sum_{\underbrace{g,c}_{x}} \underbrace{P(g|I=1)P(c|g,I=1)}_{P(x)} \underbrace{\omega(I=1,g,c,A=1)}_{f(x)} \right)$$

[from (16)]

$$= \lim_{N\to\infty} \sum_{g,c} P(g|I=1)P(c|g,I=1)P(I=1)P(A=1|g,c)$$

$$= \lim_{N\to\infty} \sum_{g,c} P(I=1,g,c,A=1)$$

[def'n of Bayes Net]

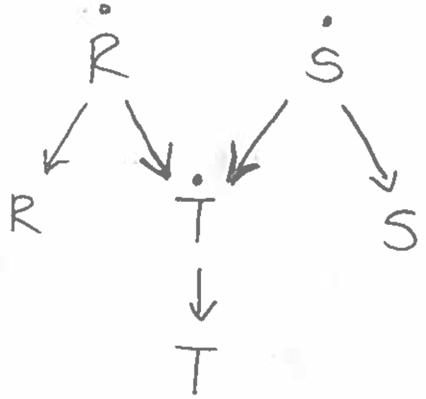
$$= \lim_{N\to\infty} P(I=1, A=1)$$

$$= P(I=1, A=1)$$

This sampling method is called likelihood weighting (and is an instance of something called importance sampling).

INFERENCE BY SAMPLING

18 While likelihood weighting is usually more efficient than rejection sampling, it's still not bulletproof. Consider again the blood types network.



Imagine that B genes are extremely rare, e.g.

$\dot{R}$	$P(\dot{R})$
AA	.2
AB	.00001
AO	.29997
BB	.00001
BO	.00001
OO	.5

19 Now let's do likelihood weighting for $P(R=AB, T=B)$:

variable	set	sample	weight
$\dot{R}$		$\dot{R}=AO$	1
$\dot{S}$		$\dot{S}=OO$	1
R	R=AB		$P(R=AB \mid \dot{R}=AO)=0$
S		S=O	1
$\dot{T}$		$\dot{T}=AO$	1
T	T=B		$P(T=B \mid \dot{T}=AO)=0$
			0

INFERENCE BY SAMPLING

(20) Our sample has weight 0, so we are effectively throwing it away (just like with rejection sampling). This is because our sampling process is still uninfluenced by the evidence — likelihood weighting just does some corrections after the fact.

But for the blood type network, it's too little, too late.

(21) The symptom: we waste a lot of samples
The root problem: our sampling process is not influenced by the evidence

Rejection sampling doesn't address the symptom. Likelihood sampling addresses the symptom, but not the problem.

Is there a sampling technique that addresses the root problem directly?